Prediksi Sektor Industri Saham BEI dengan Metode Linear Discriminant Analysis berdasarkan laporan keuangan

Muhammad Rafid Mustaghfirin
Fakultas Informatika,
Universitas Telkom, Bandung
mustrafid@students.telkomuniversity.ac.id

Deni Saepudin
Fakultas Informatika,
Universitas Telkom, Bandung
denisaeupudin@telkomuniversity.ac.id

Abstrak

Pengelompokan perusahaan berdasarkan sektor industri merupakan aspek penting dalam analisis investasi, namun klasifikasi yang dilakukan secara manual di Bursa Efek Indonesia (BEI) masih kurang optimal dalam memanfaatkan data laporan keuangan sebagai dasar prediksi. Oleh karena itu, penelitian ini mengembangkan model prediksi sektor industri berdasarkan laporan keuangan menggunakan metode Linear Discriminant Analysis (LDA) dan membandingkannya dengan Extreme Gradient Boosting (XGBoost). Data laporan keuangan perusahaan yang terdaftar di BEI dalam rentang 2010–2022 digunakan sebagai dataset utama, dengan proses pre-processing, normalisasi, dan oversampling menggunakan Borderline-SMOTE untuk mengatasi ketidakseimbangan kelas. Evaluasi model dilakukan dengan metrik accuracy, precision, recall, dan F1-score, serta dilakukan analisis fitur menggunakan Permutation Importance untuk menentukan variabel yang paling berpengaruh. Hasil penelitian menunjukkan bahwa metode LDA memiliki akurasi 27,51%, sedangkan XGBoost mencapai 63,87%, yang menunjukkan bahwa pendekatan non-linear XGBoost lebih unggul dalam mengklasifikasikan sektor industri berdasarkan laporan keuangan. Selain itu, fitur total aset, total pendapatan, dan inventaris ditemukan sebagai variabel paling berpengaruh dalam prediksi. Penelitian ini memberikan kontribusi dalam pengembangan metode otomatis untuk klasifikasi sektor industri, yang dapat digunakan oleh investor dan analis dalam mendukung pengambilan keputusan investasi yang lebih akurat.

Kata Kunci: bursa efek indonesia, linear discriminant analysis, xgboost, laporan keuangan, prediksi sektor industri

Abstract

Classifying companies based on their industry sector is crucial for investment analysis. However, the current manual classification used by the Indonesia Stock Exchange (IDX) is not optimal in leveraging financial statement data for industry sector prediction. Therefore, this study develops a predictive model using Linear Discriminant Analysis (LDA) and compares it with Extreme Gradient Boosting (XGBoost). The dataset consists of financial reports from companies listed on IDX between 2010 and 2022, which undergo pre-processing, normalization, and oversampling using Borderline-SMOTE to address class imbalance. Model evaluation is conducted using accuracy, precision, recall, and F1 score, while Permutation Importance is applied to identify the most influential financial features. The results show that LDA achieves an accuracy of 27.51%, whereas XGBoost outperforms it with an accuracy of 63.87%, indicating that the non-linear XGBoost approach is more effective for industry sector classification. Additionally, total assets, total revenue, and inventory are identified as the most significant factors in predicting industry sectors. This study contributes to the development of an automated classification method based on financial reports, which can assist investors and analysts in making more accurate investment decisions.

Keywords: Indonesia stock exchange, linear discriminant analysis, xgboost, financial statements, industry sector prediction

1. PENDAHULUAN

Latar Belakang

Dalam rangka menangani permasalahan yang terkait dengan prediksi sektor industri perusahaan berdasarkan data laporan keuangan, akan dibahas masalah pengelompokan perusahaan ke dalam sektor industri yang tepat berdasarkan data laporan keuangan. Tugas akhir ini mengambil pendekatan dengan menggunakan metode Linear Discriminant Analysis (LDA). Di Bursa Efek Indonesia, terdapat 12 sektor industri yang telah diklasifikasikan oleh Indonesia Stock Exchange Industrial Classification (IDX-IC).

Sejumlah penelitian yang relevan dilakukan sebelum penelitian ini dilakukan oleh Hans Van Der Heijden. Pada tahun 2022, Hans Van Der Heijden mengevaluasi proyeksi kinerja saham perusahaan di sektor industri Amerika Utara (NAICS) dengan menggunakan data laporan keuangan masing-masing entitas. Dalam penelitiannya, studi ini menerapkan metode Linear Discriminant Analysis (LDA) dan metode Random Forest untuk membandingkan pendekatan non-linear dan linear dalam prediksi. Temuannya mengindikasikan bahwa metode Random Forest, sebagai pendekatan non-linear, menunjukkan tingkat akurasi yang lebih tinggi dibandingkan dengan metode LDA. Dengan demikian, dapat disimpulkan bahwa Random Forest lebih efektif dalam memprediksi sektor industri dibandingkan dengan LDA [1] .

Selain Random Forest, metode XGBoost juga merupakan salah satu pendekatan non-linear yang sering digunakan dalam klasifikasi dan prediksi berbasis laporan keuangan. XGBoost merupakan algoritma berbasis pohon keputusan yang dioptimalkan untuk meningkatkan akurasi dan efisiensi perhitungan dalam permasalahan klasifikasi. Beberapa penelitian menunjukkan bahwa XGBoost memiliki performa yang baik dalam menangani data keuangan. Studi oleh Chen dan Guestrin (2016) menunjukkan bahwa XGBoost mampu memberikan hasil klasifikasi yang lebih akurat dibandingkan dengan metode lain dalam berbagai kasus, termasuk prediksi keuangan [2]. Sementara itu, penelitian 1 lain oleh Li et al. (2021) mengungkapkan bahwa XGBoost lebih unggul

dibandingkan model tradisional dalam memprediksi kebangkrutan perusahaan berdasarkan laporan keuangan [3].

Sejumlah penelitian juga telah dilakukan untuk mengevaluasi penerapan metode klasifikasi berbasis laporan keuangan, salah satunya adalah penelitian mengenai penggunaan Multiple Discriminant Analysis (MDA) dalam memprediksi financial distress perusahaan. MDA, yang merupakan bentuk umum dari Linear Discriminant Analysis (LDA), telah digunakan dalam analisis perusahaan sektor industri barang konsumsi yang terdaftar di Bursa Efek Indonesia. Hasil penelitian menunjukkan bahwa metode ini mampu membedakan perusahaan yang mengalami kesulitan keuangan dan yang tidak, sehingga dapat menjadi pendekatan yang efektif dalam analisis laporan keuangan untuk prediksi kinerja perusahaan [4].

Dalam tugas akhir ini, penelitian akan memfokuskan diri pada penggunaan algoritma LDA untuk mengatasi masalah prediksi sektor industri perusahaan. Dengan memanfaatkan fitur-fitur dalam laporan keuangan sebagai prediktor, algoritma ini diharapkan dapat memberikan hasil prediksi yang akurat dan dapat diinterpretasikan. Selain itu, untuk membandingkan efektivitas pendekatan linear dan non-linear, penelitian ini juga akan menambahkan metode XGBoost sebagai pembanding. Dengan adanya perbandingan ini, diharapkan dapat diperoleh pemahaman yang lebih dalam mengenai efektivitas kedua algoritma dalam menyelesaikan permasalahan klasifikasi sektor industri berdasarkan data laporan keuangan.

LDA dipilih dalam penelitian ini karena berfokus pada pendekatan linear, mengingat belum adanya penelitian sebelumnya yang secara spesifik mengangkat topik ini berdasarkan laporan keuangan. Pendekatan linear dalam klasifikasi sektor industri masih jarang dieksplorasi dalam literatur, sehingga penelitian ini diharapkan dapat memberikan kontribusi baru dalam memahami efektivitas metode LDA dalam konteks tersebut. Selain itu, LDA tetap memiliki keunggulan dalam interpretabilitas dan efisiensi komputasi, menjadikannya pilihan yang relevan dalam analisis laporan keuangan.

Topik dan Batasannya

Dalam penelitian ini, topik yang dianalisis oleh penulis adalah bagaimana membangun model prediksi menggunakan metode Linear Discriminant Analysis (LDA) untuk setiap sektor industri dengan membandingkan performa model LDA dengan metode pembanding yaitu Algoritma model XGBoost. Sebelumnya data akan diolah menggunakan metode oversampling untuk penyeimbangan data pada data yang tidak seimbang. Batasan dalam penelitian ini adalah keterbatasan data laporan keuangan yang tersedia di Bursa Efek Indonesia (BEI), di mana data laporan keuangan hanya berbentuk dokumen sehingga perlu dilakukan input manual untuk data dari masing-masing sektor industri.

Tujuan

Berdasarkan rumusan masalah yang dijelaskan pada bagian sebelumnya, maka penelitian ini akan memiliki beberapa tujuan diantaranya adalah :

1. Mengimplementasikan algoritma Linear Discriminant Analysis (LDA) dan XGBoost untuk memprediksi sektor industri perusahaan berdasrkan laporan keuangannya.
2. Menganalisis perbandingan dari dua pendekatan algoritma yang berbeda yaitu metode algoritma LDA dan XGBoost dalam memprediksi sektor industri perusahaan berdasarkan laporan keuangan.
3. Mengidentifikasi fitur-fitur yang memiliki dampak paling signifikan dalam meningkatkan akurasi prediksi sektor industri perusahaan menggunakan algoritma LDA dan XGBoost.

TABEL 1
Tabel keterkaitan antara tujuan, pengujian dan kesimpulan

no.	Tujuan	Pengujian	Kesimpulan
1	Mengimplementasikan algoritma Linear Discriminant Analysis (LDA) dan XGBoost untuk memprediksi sektor industri perusahaan berdasrkan laporan keuangannya.	Pengujian Model LDA dan XGBoost dengan mengkombinasikan fitur historis dan fundamental dilakukan menggunakan metrik akurasi dan F1 Score	Model prediksi LDA dan XGBoost berhasil dikembangkan
2	Menganalisis perbandingan dari dua pendekatan algoritma yang berbeda dalam memprediksi sektor industri perusahaan berdasarkan laporan keuangan	Menganalisis perbandingan dari hasil implementasi pengujian model LDA dan XGBoost	Hasil akurasi model yang paling tinggi dianggap model yang berhasil untuk memprediksi sektor industri saham perusahaan berdasarkan laporan keuangan
3	Mengidentifikasi fitur-fitur yang memiliki dampak paling signifikan dalam meningkatkan akurasi prediksi sektor industri perusahaan menggunakan algoritma LDA.	Pengujian Model LDA dengan memanfaatkan metode Permutation Importance untuk melihat fitur - fitur yang paling berpengaruh dalam memprediksi sektor industri	Fitur fitur yang paling berpengaruh dipilih berdasarkan performa model prediksi.

2. KAJIAN TEORI

2.1 Penelitian Terkait

Suatu Penelitian yang relevan dilakukan sebelum penelitian ini dilakukan oleh Peter Boedeker and Nathan T. (2019) dalam penelitian mereka mengenai penerapan Linear Discriminant Analysis (LDA) untuk memprediksi Group Membership. Mereka menyoroti keefektifan LDA dalam kasus di mana jumlah prediktor kurang dari setengah dari jumlah kasus, terdapat korelasi yang signifikan antar prediktor, dan asumsi model ini hampir terpenuhi. LDA dianggap lebih unggul dibandingkan dengan metode klasifikasi lain seperti regresi logistik dan logistik multinomial, terutama dalam situasi ketidaksesuaian spesifikasi kelompok referensi atau saat suatu kondisi 8 yang dapat dipisahkan [5].

Penelitian yang dilakukan Noviyanti Santoso and Wahyu Wibowo pada tahun 2018 menyoroti keberhasilan Model Linear Discriminant Analysis (LDA) dan Support Vector Machine (SVM) dalam memprediksi kesulitan keuangan. Hasil prediksi keduanya dinilai memuaskan, namun model SVM dengan periode dinamis k = 1 dan metode hybrid, terutama dengan seleksi variabel secara stepwise, terbukti menjadi model terbaik di antara empat kombinasi model dengan mencapai nilai Area Under the Curve (AUC) dan akurasi maksimum. Pentingnya pemilihan variabel ditekankan sebagai faktor peningkatan signifikan dalam akurasi prediksi baik pada model LDA maupun SVM [6].

Suatu Penelitian yang dilakukan oleh Nur Febrianti Bakri pada tahun 2024, dengan tujuan memprediksi kondisi financial distress pada perusahaan sektor industri di Indonesia menggunakan metode Extreme Gradient Boosting (XGBoost) berdasarkan data laporan keuangan dari tahun 2005 hingga 2018. Hasil penelitian menunjukkan bahwa model XGBoost tanpa synthetic features generation memiliki akurasi tertinggi sebesar 95,86%, AUC 94,22%, dan F1-score 97,08% pada time window size 0, namun performanya menurun seiring peningkatan ukuran time window. Dengan synthetic features generation dan seleksi variabel, performa model meningkat, terutama pada size 1 dengan akurasi 88,28%, AUC 85,31%, dan F1-score 91,71%, meskipun pada ukuran yang lebih besar performanya menurun. Dibandingkan dengan metode lain seperti Generalized Extreme Value Regression (GEVR), Logistic Regression, dan Support Vector Machine (SVM), model XGBoost dengan synthetic features generation dan seleksi variabel terbukti lebih unggul dalam memprediksi financial distress. Secara keseluruhan, penelitian ini menyimpulkan bahwa XGBoost merupakan metode yang efektif dalam memprediksi financial distress pada perusahaan sektor industri di Indonesia[7].

Suatu studi pada tahun 2019 oleh Rian Dolphin , Barry Smyth, Ruihai Dong menyoroti pentingnya laporan keuangan karena mengandung informasi relevan tentang arus kas perusahaan yang dapat diprediksi. Studi ini menunjukkan bahwa model machine learning dapat menghasilkan prediksi reaksi pasar yang lebih akurat dibandingkan dengan metode tradisional. Penelitian tersebut menegaskan bahwa laporan keuangan menyediakan data penting yang mencerminkan kondisi keuangan perusahaan, seperti arus kas, pendapatan, dan pengeluaran. Dengan menerapkan algoritma machine learning pada data ini, model dapat mengidentifikasi pola dan tren yang mungkin tidak terdeteksi oleh metode analisis tradisional. Hal ini memungkinkan prediksi yang lebih akurat terkait bagaimana pasar akan bereaksi terhadap informasi keuangan tertentu[8].

2.2 Laporan Keuangan

Laporan keuangan adalah catatan informasi keuangan dari suatu perusahaan pada suatu periode akuntansi [9]. Keberadaan laporan keuangan dapat digunakan untuk menggambarkan kinerja perusahaan khususnya dalam bidang keuangan [10]. Susunan laporan keuangan terbagi menjadi laporan posisi keuangan, laporan laba rugi, laporan arus kas dan laporan perubahan modal dan catatan atas laporan keuangan [11]. Menurut Werner R. Murhadi (2019) laporan keuangan merupakanbentuk bahasa bisnis. Laporan keuangan memberikan data yang terolah kepada pengguna tentang posisi keuangan perusahaan. Memahami laporan keuangan perusahaan memungkinkan pemangku kepentingan yang berbeda untuk memahami posisi keuangan perusahaan [12].

2.3 Linear Discriminant Analysis

Model Linear Discriminant Analysis (LDA) digunakan untuk ekstraksi fitur kedalam kelas-kelas yang berbeda. model Linear Discriminant Analysis (LDA) mampu memisahkan antarkelas menjadi lebih terpisah dengan memaksimalkan nilai between-class scatter dan meminimalkan within-class scatter. Pada ekstraksi ciri menggunakan Linear Discriminant Analysis (LDA) dataset lokasinya tetap namun kelas yang dibentuk menjadi lebih terpisah sehingga dapat menyebabkan jarak antar kelas menjadi lebih besar, sedangkan jarak antar data pelatihan dalam satu kelas akan menjadi lebih kecil (Widyati dkk, 2021) [13].

Menurut Raschka dan Mirjalili (2019), metode linear discriminant analysis ini dapat diringkas ke dalam tujuh langkah berikut:

1. Standarisasikan himpunan data awal berdimensi

$$X' = \frac{X - \mu X}{\sigma X} \quad (2.1)$$

Di mana X' adalah data awal, μX adalah rata-rata dari data, dan σX adalah deviasi standar.

2. Hitung vektor rata-rata (mean) dimensi untuk setiap kelas pada data.

$$\mu i = \frac{1}{Ni} \sum_{xj \in Ci} xj \quad (2.2)$$

Di mana μi adalah rata-rata untuk kelas. Ci, dan Ni adalah jumlah sampel dalam kelas Ci

3. Buat matriks persebaran antar-kelas (between-class scatter matrix ; S_B) dan matriks persebaran dalam-kelas (within-class scatter matrix ; S_W).

4. Hitung kumpulan eigenvector beserta eigenvalue yang berkaitan dari matriks.

$$S_W^{-1} S_B = \lambda \boldsymbol{\upsilon} \quad (2.3)$$

Di mana λ adalah eigenvalue dan $\boldsymbol{\upsilon}$ adalah eigenvector.

5. Urutkan kumpulan eigenvalue yang ada dari nilai terbesar ke terkecil (descending).

Setelah menghitung eigenvalue dan eigenvector, urutkan $\lambda 1 \geq \lambda 2 \geq \cdots \geq \lambda k$ dan pilih eigenvector
terkait dengan eigenvalue terbesar untuk membentuk matriks transformasi W.
6. Ambil sebanyak λk eigenvector dengan eigenvalue terbesar untuk membentuk matriks
transformasi W yang berdimensi, di mana setiap eigenvector mewakili satu kolom.
7. Gunakan matriks transformasi W tersebut untuk mentransformasikan matriks data awal X berdimensi yang diteliti ke dalam matriks fitur baru berdimensi k

$$X_{transformed} = XW \quad (2.4)$$

Dimana $X_{transformed} = XW$ adalah data yang sudah ditransformasikan ke dalam ruang fitur baru berdimensi k [14].
Teknik Linear Discriminant Analysis bertujuan untuk memproyeksikan matriks data asli ke dalam ruang berdimensi lebih rendah. Untuk mencapai tujuan ini, ada tiga langkah yang perlu dilakukan Langkah pertama adalah menghitung pemisahan antara kelas-kelas yang berbeda (yaitu, jarak antara rerata kelas yang berbeda), yang disebut sebagai the between-class variance atau between-class matrix. Langkah kedua adalah menghitung jarak antara rerata dan sampel-sampel setiap kelas, yang disebut sebagai within-class variance atau within class matrix. Langkah ketiga adalah membangun ruang berdimensi lebih rendah yang memaksimalkan variasi antar-kelas dan meminimalkan variasi dalam-kelas [15].

a) Between-class matrix

Matriks between-class variance (variansi antar kelas) adalah sebuah matriks yang digunakan dalam Analisis Diskriminan Linear (Linear Discriminant Analysis atau LDA) untuk mengukur sejauh mana pusat-pusat massa atau rerata dari kelas kelas yang berbeda tersebar satu sama lain. Tujuannya adalah untuk menemukan 21 proyeksi linier yang memaksimalkan jarak antara rerata kelas, sehingga memperkuat pemisahan antara kelas. dalam konteks pengklasifikasian laporan keuangan (misalnya, sektor industri, jenis bisnis, atau tingkat risiko), between-class variance dapat membantu membedakan karakteristik keuangan antara kategori-kategori ini. Ini dapat berguna untuk memahami perbedaan penting dalam kinerja keuangan antar kelas tersebut. [16].

$$S_B = \sum_{i=1}^{k} n_i (m_i - m)(m_i - m)^T \quad (2.5)$$

Di mana S_B adalah rata-rata keseluruhan dari seluruh data, k adalah jumlah kelas, m_i adalah rata-rata kelas i, dan n_i adalah jumlah elemen atau sampel dalam kelas ke-i. m adalah vektor rerata keseluruhan .

b) Within-class matrix

Matriks dalam kelas (within-class matrix atau S_W) adalah komponen penting dalam Analisis Diskriminan Linear (Linear Discriminant Analysis atau LDA). Matriks ini digunakan untuk mengukur variasi atau perbedaan antara sampel sampel dalam satu kelas. Dalam konteks LDA, tujuan dari $S_W$adalah untuk meminimalkan variasi dalam satu kelas dan, pada gilirannya, meningkatkan kohesivitas atau kesamaan antar sampel dalam kelas yang sama [17]

$$S_W = \sum_{i=1}^{c} S_i \sum_{j=1}^{Mi} (x_j - \mu_i)(x_j - \mu_i)^T \quad (2.6)$$

Dimana $\sum_{j=1}^{Mi} (x_j - \mu_i)(x_j - \mu_i)^T$ merupakan matriks covarian. Mi sebagai jumlah sampel dalam kelas ke-i, dan x_j yang merepresentasikan vektor fitur dari sampel ke- j dalam kelas ke- i. Selain itu, μ_i adalah rata-rata vektor fitur dari semua sampel dalam kelas ke- i, sedangkan $(x_j - \mu_i)$ menunjukkan deviasi setiap sampel terhadap rata-rata kelasnya. Terakhir, $(x_j - \mu_i)^T$ merupakan transpos dari deviasi vektor tersebut, yang digunakan untuk menghitung matriks kovarians dalam kelas guna mengukur variasi antar sampel dalam kelas yang sama [18].

2.4 Oversampling Borderline-Smote

Pada Tahap ini, teknik oversampling Borderline-SMOTE (Synthetic Minority Over-sampling Technique) digunakan untuk menangani masalah ketidakseimbangan kelas dalam dataset. Borderline-SMOTE adalah salah satu varian dari teknik Synthetic Minority Oversampling Technique (SMOTE) yang digunakan untuk mengatasi masalah ketidakseimbangan data (class imbalance). Ketidakseimbangan data terjadi ketika jumlah sampel dari kelas minoritas jauh lebih kecil dibandingkan dengan kelas mayoritas, yang dapat menyebabkan model pembelajaran mesin tidak mampu mengenali kelas minoritas dengan baik [19].
Terdapat perbedaan antara Metode SMOTE biasa dengan Borderline-SMOTE. SMOTE (Synthetic Minority Over-sampling Technique) adalah teknik oversampling yang menghasilkan sampel sintetis berdasarkan interpolasi antara sampel minoritas yang ada [20]. Sementara itu, Borderline-SMOTE, yang dikembangkan oleh Han et al. (2005), merupakan variasi SMOTE yang lebih fokus pada sampel minoritas yang berada di dekat batas keputusan (borderline), di mana kemungkinan mis-klasifikasi lebih tinggi. Borderline-SMOTE secara khusus memilih sampel minoritas yang memiliki tetangga mayoritas lebih banyak, sehingga sampel sintetis yang dihasilkan lebih informatif untuk memperbaiki pemisahan kelas dibandingkan SMOTE biasa yang melakukan oversampling secara acak. Metode ini bekerja dengan terlebih dahulu mengidentifikasi contoh minoritas yang berada di area kritis, yaitu di dekat perbatasan dengan kelas mayoritas, menggunakan k-nearest neighbors (k-NN). Jika sebagian besar dari k tetangga terdekat suatu sampel minoritas berasal dari kelas mayoritas, maka sampel tersebut dikategorikan sebagai borderline dan diprioritaskan untuk oversampling. Setelah identifikasi ini, Borderline-SMOTE menghasilkan sampel sintetis dengan cara interpolasi antara contoh borderline minoritas dan tetangga minoritas terdekatnya, serupa dengan SMOTE, tetapi dengan lebih menekankan pada contoh yang berisiko tinggi untuk salah klasifikasi. Dengan pendekatan ini, Borderline-SMOTE bertujuan untuk memperbaiki representasi kelas minoritas di area yang paling menentukan dalam pembelajaran mesin, sehingga meningkatkan akurasi model klasifikasi

dalam mendeteksi dan mengenali pola data dari kelas yang kurang terwakili [21].

2.5 Seleksi Feature menggunakan Permutation Importance

Seleksi Feature menggunakan Permutation Importance Seleksi fitur merupakan langkah penting dalam proses analisis data untuk meningkatkan akurasi model serta mengurangi kompleksitas komputasi. Dalam penelitian ini, seleksi fitur dilakukan menggunakan Permutation Importance, yaitu teknik yang digunakan untuk mengidentifikasi fitur-fitur yang paling berpengaruh dalam Prediksi Sektor Industri Saham Bursa Efek Indonesia (BEI) dengan Metode Linear Discriminant Analysis (LDA) berdasarkan laporan keuangan.

Permutation Importance merupakan metode yang digunakan untuk mengukur kontribusi setiap fitur terhadap kinerja model dengan cara mengacak (permutasi) nilai suatu fitur secara independen, kemudian mengukur perubahan performa model setelahnya. Jika hasil prediksi model mengalami penurunan signifikan setelah permutasi, maka fitur tersebut dianggap penting. Sebaliknya, jika tidak ada perubahan berarti, fitur tersebut kemungkinan kurang berpengaruh dalam proses klasifikasi [22].

Dalam penelitian ini, dataset yang digunakan berasal dari laporan keuangan perusahaan yang terdaftar di Bursa Efek Indonesia (BEI). Berbagai indikator keuangan, dianalisis untuk menentukan sektor industri suatu saham. Dengan menggunakan Permutation Importance, fitur-fitur utama yang berkontribusi dalam pemisahan sektor industri oleh LDA dapat diidentifikasi. Hasil seleksi fitur ini akan membantu dalam penyederhanaan model, mengurangi dimensi data, dan meningkatkan interpretabilitas hasil klasifikasi. Dengan demikian, model dapat lebih efektif dalam memprediksi sektor industri saham berdasarkan laporan keuangan, yang pada akhirnya dapat mendukung pengambilan keputusan investasi yang lebih akurat.

2.5 Extreme Gradient Boosting (XGBoost)

XGBoost (EXtreme Gradient Boosting) adalah sebuah algoritma machine learning yang digunakan untuk memprediksi nilai numerik atau kategori dari suatu data. Algoritma ini termasuk dalam kategori ensemble learning, yang menggabungkan beberapa model machine learning untuk meningkatkan akurasi prediksi. XGBoost menggunakan teknik gradient boosting untuk menghasilkan model yang lebih akurat dan efisien [23].

XGBoost membuat sebuah sistem pohon boosting yang efisien dan scalable yang dapat digunakan untuk memprediksi nilai numerik atau kategori dari suatu data. Algoritma ini menggunakan teknik gradient boosting untuk meningkatkan akurasi prediksi dan mengatasi masalah overfitting. XGBoost juga dilengkapi dengan fitur-fitur seperti regularisasi, parallel processing, dan handling missing values [24]. XGBoost menggunakan teknik Gradient Boosting, yang secara matematis didasarkan pada optimasi fungsi loss menggunakan turunan gradien. Berikut adalah rumus utama yang digunakan dalam XGBoost :

a) Model Ensemble

XGBoost membangun model prediksi dengan menggabungkan beberapa pohon keputusan :

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \quad (2.7)$$

di mana:

- $\hat{y}_i$ adalah prediksi akhir untuk sampel ke-i,
- K adalah jumlah total pohon,
- f_k adalah pohon keputusan ke- k,
- x_i adalah fitur dari sampel ke- i.

b) Fungsi Objektif dalam XGBoost

Fungsi objektif dalam XGBoost mengoptimalkan dua komponen: fungsi loss (L) dan regularisasi (Ω):

$$Obj(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (2.8)$$

di mana:

- $L(y_i, \hat{y}_i)$ adalah fungsi loss untuk mengukur seberapa baik model memprediks

i nilai target $\hat{y}_i$

- $\Omega(f_k)$adalah istilah regularisasi untuk mengontrol kompleksitas model, biasanya berbentuk:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j} w_j^2 \quad (2.9)$$

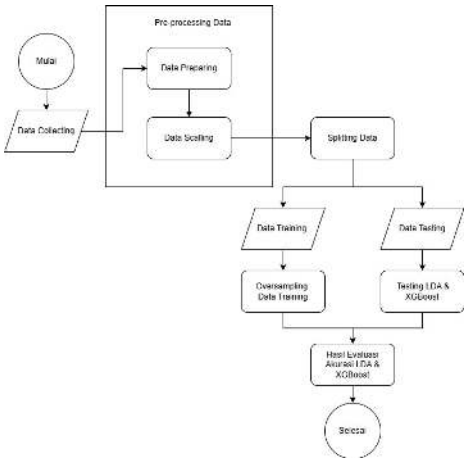
dengan:

- T adalah jumlah daun dalam pohon keputusan,
- w_j adalah bobot pada setiap daun,
- γ dan λ adalah parameter regularisasi.

3. METODE

3.1. Desain Sistem

Alur sistem dimulai dengan pengumpulan data (Data Collecting), kemudian melalui tahap Pre-processing Data yang mencakup Data Preparing (pembersihan data) dan Data Scaling (normalisasi atau standarisasi fitur). Normalisasi atau standarisasi data agar setiap fitur



GAMBAR 3.1.
Desain Sistem

memiliki skala yang seragam, sehingga model bekerja lebih optimal. Data yang telah diproses kemudian dibagi menjadi dua bagian utama Data Training dan Data Testing. Pada tahap pelatihan, dilakukan oversampling menggunakan Borderline-SMOTE, yang berfokus pada pembuatan sampel sintetis dari kelas minoritas yang berada di sekitar batas keputusan (decisionboundary), sehingga model lebih mampu mengenali pola dalam kelas minoritas tanpa menyebabkan overfitting. Setelah itu, dilakukan

pengujian menggunakan Linear Discriminant Analysis (LDA), yang membantu dalam reduksi dimensi untuk meningkatkan separabilitas kelas, serta XGBoost, algoritma ensemble berbasis pohon yang mengoptimalkan gradient boosting untuk meningkatkan akurasi prediksi. Hasil dari kedua model dibandingkan berdasarkan metrik evaluasi seperti akurasi, precision, recall, dan F1-score untuk menentukan model terbaik. Proses ini berakhir setelah hasil evaluasi dianalisis dan mebandingkan kinerja dari kedua model.

3.2. Pengumpulan Data

Dalam Bagian data ini dibagi menjadi tiga bagian. Bagian pertama akan membahas secara spesifik kumpulan data. Bagian kedua akan membahas tanggungan variabel, yang dalam literatur pembelajaran mesin disebut sebagai "target". Bagian selanjutnya akan membahas tentang variabel bebas yang disebut "features"

TABEL 3.1.
Mapping Sektor Industri IDX

Target Class Code	IDX – IC Code	Description	Acronym
0	A	Energy	ENE
1	B	Basic Materials	BAS
2	C	Industrials	IND
3	D	Consumer Non-Cyclicals	CNC
4	E	Consumer Cyclicals	COC
5	F	Healthcare	HEA
6	G	Financials	FIN
7	H	Properties & Real Estate	PRO
8	I	Technology	TEC
9	J	Infrastructures	INF
10	K	Transportation & Logistic	TRA

Sumber data ini diambil dari Bursa Efek Indonesia (BEI) yang terbaru, yaitu IDX-IC yang terdiri dari 11 Sektor. Bursa Efek Indonesia (BEI) menggunakan IDX Industrial Classification atau IDX-IC untuk mengklasifikasikan Perusahaan yang terdaftar. Penentuan sektor, sub-sektor, industri, atau sub-industri didasarkan pada eksposur pasar. BEI memiliki hak untuk menetapkan klasifikasi Perusahaan yang terdaftar berdasarkan penilaian dan justifikasi yang dilakukan oleh BEI.

TABEL 3.2.
Jumlah Data Seluruh Perusahaan

Target Class Code	Sector	Jumlah Perusahaan	Presentase
0	A	437	10.18%
1	B	634	14.76%
2	C	480	11.18%
3	D	716	16.67%
4	E	715	16.65%
5	F	155	3.61%
6	G	85	1.98%
7	H	412	9.59%
8	I	96	2.24%
9	J	375	8.73%
10	K	185	4.40%
Total		4290	100%

Dari data yang sudah diperoleh Ketidakseimbangan data terlihat jelas dari perbedaan signifikan dalam jumlah perusahaan di setiap sektor. Beberapa sektor memiliki jumlah perusahaan yang sangat tinggi, sementara sektor lainnya memiliki jumlah yang relatif rendah. Perbedaan yang signifikan ini menunjukkan adanya ketidakseimbangan yang bisa berdampak pada hasil prediksi yang dilakukan nanti. Untuk mengatasi ketidakseimbangan data ini, beberapa pendekatan akan dilakukan, seperti oversampling sektor dengan jumlah perusahaan yang lebih sedikit untuk membuktikan hasil dari data yang sudah di oversampling dan dengan yang belum.

3.3. Pre-processing Data

Pada tahap pra-pemrosesan dilakukan pembersihan dengan kriteria untuk menghapus subset data yang tidak lengkap dari semua atribut (variabel) yang digunakan. Setelah dibersihkan, transformasi dilakukan dengan membuat data mengikuti distribusi normal (menormalkan data yang menjadikan rentang skalanya 0 hingga 1).

Normalisasi data dilakukan menggunakan metode Min-Max Scaler. Min-Max Scaler merupakan salah satu teknik normalisasi yang mengubah data sehingga berada dalam rentang tertentu, umumnya antara 0 hingga 1 [25]. Proses ini dilakukan dengan rumus sebagai berikut:

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (3.1)$$

Dimana :

- X adalah nilai asli dari suatu atribut
- X_{min} adalah nilai minimum dari atribut tersebut dalam dataset
- X_{max} adalah nilai maksimum dari atribut tersebut dalam dataset
- $X_{normalized}$ adalah nilai hasil normalisasi dalam rentang (0,1).

Setelah pembersihan dan transformasi, pengurangan dimensi dilakukan dilakukan pada variabel/fitur independet. Variable yang akan dipakai nanti menggunakan laporan keungan dari tahun 2010 – 2022 yang dibagi menjadi beberapa financial ratio yang sudah dibagi dibawah ini:

TABEL 3.3.
Data Feature Laporan Keuangan

Variable	Common Size Component	Financial Statement
B1	Balance Sheet	Cash
B2		Short Investment
B3		Trade Recievables
B4		Other Recievables
B5		Inventories Current Asset
B6		Others Current Asset
B7		Total Current Asset
B8		Non Current Asset
B9		Total Assets
B10		Current Liabilities
B11		Non Current Liabilities
B12		Total Liabilities
B13		Total Equity
I1	Income Statement	Total Revenue
I2		Gross Profit
I3		Income From Operations
I4		Income Before Tax
I5		Net Income For The Period
I6		Total Comprehensive Income

3.4. Pembagian Data

Pada tahap ini, data dibagi menjadi beberapa bagian untuk keperluan pelatihan, validasi, dan pengujian model. Pembagian data yang tepat sangat penting dalam memastikan bahwa model yang dibangun memiliki performa yang baik dan dapat digeneralisasikan dengan baik ke data yang belum pernah dilihat sebelumnya. Berikut tabel pembagian data training dan test yang akan digunakan untuk model prediksi ini :

TABEL 3.3.
Data Splitting

Data Training	80%
Data Testing	20%

Merujuk pada Tabel 3.4 data di split atau data dibagi dengan komposisi data 80% untuk data training dan 20% untuk data testing. Hal ini digunakan untuk menghindari overfitting dan memastikan evaluasi objektif terhadap model prediksi.

3.5. Skenario Pengujian

Pengujian dilakukan untuk memastikan bahwa akurasi dari prediksi bisa mendapatkan hasil yang sesuai, dengan adanya skenario pengujian ini dibuat karena adanya improvement terkait hasil prediksi sebelum oversampling dan sesudah oversampling. Berikut adalah beberapa skenario pengujian yang akan dilakukan:

NO	Skenario Pengujian	Hasil yang diharapkan	Hasil Pengujian	Ket
`				LDA Model
1	Melakukan prediksi model LDA dengan data yang sudah di oversampling			
				XGBoost Model
2.	Melakukan prediksi model XGBoost dengan data yang sudah di oversampling			

Skenario ini dilakukan berdasarkan pengujian prediksi Model LDA dan XGBoost yang akan dilakukan sesuai yang dimana dilakukan sesudah oversampling

3.6. Evaluasi Performansi Model

3.6.1. Confusion Matrix

Confusion matrix adalah salah satu tools analitik prediktif yang menampilkan dan membandingkan nilai aktual atau nilai sebenarnya dengan nilai hasil prediksi model yang dapat digunakan untuk menghasilkan metrik evaluasi seperti Accuracy (akurasi), Precision, Recall, dan F1-Score atau F Measure. [26]

Ada empat nilai yang dihasilkan di dalam tabel confusion matrix, di antaranya True Positive (TP), False Positive (FP), False Negative (FN), dan True Negative (TN). Ilustrasi tabel confusion matrix dapat dilihat pada gambar berikut :

GAMBAR 3.2.
Ilustrasi tabel confusion matrix

- True Positive (TP) : Jumlah data yang bernilai Positif dan diprediksi benar sebagai Positif
- False Positive (FP) : Jumlah data yang bernilai Negatif tetapi diprediksi sebagai Positif.
- False Negative (FN) : Jumlah data yang bernilai Positif tetapi diprediksi sebagai Negatif.
- True Negative (TN) : Jumlah data yang bernilai Negatif dan diprediksi benar sebagai Negatif.

3.6.2. Accuracy

Nilai akurasi didapatkan dari jumlah data bernilai positif yang diprediksi positif dan data bernilai negatif yang diprediksi negatif dibagi dengan jumlah seluruh data di dalam dataset.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3.2)$$

3.6.3. Precision

Precision adalah peluang kasus yang diprediksi positif yang pada kenyataannya termasuk kasus kategori positif.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3.3)$$

3.6.4. F1-Score

Nilai F1-Score atau dikenal juga dengan nama F-Measure didapatkan dari hasil Precision dan Recall antara kategori hasil prediksi dengan kategori sebenarnya.

$$F1 = \frac{2 \text{ x Recall x Precision}}{\text{Recall x Precision}} \quad (3.4)$$

[26].

4. HASIL DAN PEMBAHASAN

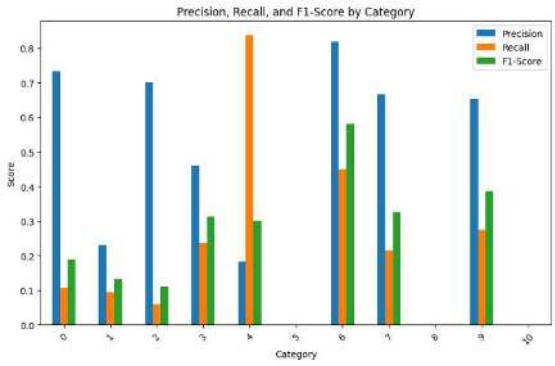
4.1 Skenario Pengujian tanpa oversampling

Total data pelaporan yang digunakan adalah 4.294 data laporan keuangan yang diambil dari hasil penyaringan laporan keuangan sektoral dari Bursa Efek Indonesia (BEI). Fase ini terjadi sebelum oversampling. Hasil dari oversampling adalah:

TABEL 4.1.
Jumlah Kelas Sebelum Oversampling

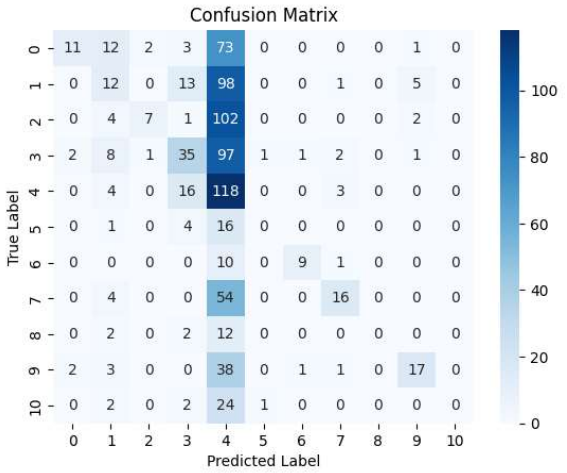
Kelas	Jumlah kelas data training sebelum oversampling
0	335
1	505
2	363
3	568
4	574
5	134
6	65
7	338
8	80
9	313
10	160

Seperti yang terlihat di tabel diatas adalah jumlah data training per-kelas yang belum dioversampling, terlihat juga bahwa jumlah pengelompokkan data yang sangat tidak seimbang hal ini bisa menyebabkan bias dalam model prediksi, di mana kelas yang memiliki data lebih banyak cenderung mendominasi prediksi model. Pada pengujian klasifikasi metode Linear Discriminant Analysis yang belum di oversampling agar terlihat perbandingan antara yang belum di oversampling untuk hasil akurasi LDA dapat dilihat pada gambar berikut:



GAMBAR 4.1.
Grafik Akurasi Dalam Model LDA Tanpa Oversampling

Laporan klasifikasi ini menunjukkan bahwa model memiliki kinerja yang rendah dengan akurasi keseluruhan hanya 26,22%. Precision tinggi pada beberapa kelas seperti 0 (73,33%), 2 (70,00%), dan 6 (81,82%), tetapi recall rendah pada hampir semua kelas, terutama kelas minoritas seperti 5, 8, dan 10, yang memiliki precision, recall, dan f1-score sebesar 0%. Ketidakseimbangan jumlah data antar kelas (misalnya, 148 data pada kelas 3 dibandingkan dengan hanya 16 pada kelas 8) menjadi salah satu penyebab utama kinerja buruk, karena model lebih condong mengenali kelas dengan data yang lebih banyak. Rata-rata makro menunjukkan precision 40,43% dan recall 20,68%, menegaskan bahwa model tidak mampu menangkap pola pada kelas-kelas minoritas,



GAMBAR 4.2.
Grafik Hasil Confusion Matrix Tanpa Oversampling

Hasil analisis data testing tanpa oversampling:

a) Model tampaknya bias terhadap kelas tertentu, khususnya kelas 4, sementara performa pada kelas lainnya, seperti 5, 8, dan 10, sangat buruk. Hal ini mungkin menunjukkan ketidakseimbangan data atau kurangnya kemampuan model untuk menangani variasi di data tersebut.

b) Bias tinggi terhadap kelas tertentu, seperti kelas 4 yang mendominasi dengan recall tinggi (0.84) tetapi precision rendah (0.18), menunjukkan bahwa banyak prediksi salah diarahkan ke kelas ini meskipun model mampu mengenali sebagian besar sampel dari kelas 4. Bias ini kemungkinan disebabkan oleh ketidakseimbangan dalam distribusi data serta model LDA yang berupaya memaksimalkan diskriminasi antar kelas tetapi gagal karena distribusi fitur antar kelas saling tumpang tindih.

c) Precision, recall, dan F1-score masing-masing berada di sekitar 0.40, 0.21, dan 0.21, menandakan performa rendah rata-rata di semua kelas tanpa memperhitungkan jumlah sampel di setiap kelas.

d) Dengan akurasi hanya 26%, model nyaris tidak lebih baik dari tebakan acak untuk dataset dengan 11 kelas. Hal ini menunjukkan bahwa model gagal menangkap pola yang bermakna dari data. Potensi penyebab performa rendah model ini adalah data yang mungkin tidak dipersiapkan dengan baik, seperti kurangnya normalisasi fitur yang dapat memengaruhi sensitivitas model LDA terhadap skala data. Selain itu, LDA mungkin tidak cocok untuk dataset ini jika hubungan antara fitur dan target bersifat non-linier, karena LDA didasarkan pada asumsi distribusi linier yang dapat menyebabkan kesulitan dalam menangkap pola yang lebih kompleks.

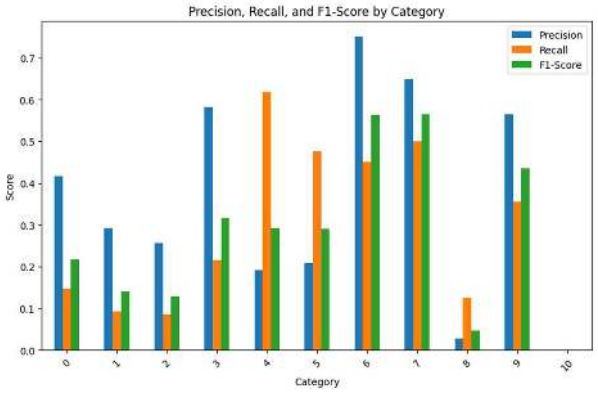
4.2 Skenario Pengujian jika menggunakan oversampling

Berdasarkan dari hasil diatas sebelum adanya oversampling dapat diketahui bahwa prediksi pada masing masing sector atau class pada penelitian ini tidak seimbang dimana class 4 memiliki presentase hasil prediksi yang lebih banyak dibanding class yang lain. Sehingga perlu dilakukan handling imbalanced data agar model yang dihasilkan tidak condong ke satu kelas. Pada penelitian ini, peneliti melakukan oversampling menggunakan Borderline-SMOTE. Adapun hasil dari oversampling tersebut adalah berikut.

TABEL 4.2.
Jumlah Kelas Setelah Oversampling

Kelas	Jumlah kelas data training sebelum oversampling	Jumlah kelas sesudah oversampling
0	335	574
1	505	574
2	363	574
3	568	574
4	574	574
5	134	574
6	65	574
7	338	574
8	80	574
9	313	574
10	160	574

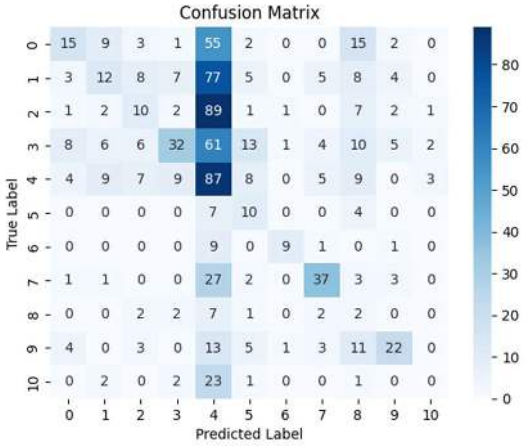
Sebelum oversampling, jumlah sampel dalam setiap kelas sangat bervariasi (misalnya, kelas 6 hanya memiliki 65 data, sedangkan kelas 4 memiliki 574 data). Setelah menerapkan Borderline-SMOTE, jumlah data dalam semua kelas disamakan menjadi 574. Teknik ini bertujuan untuk meningkatkan jumlah sampel pada kelas minoritas agar seimbang dengan kelas mayoritas, yang pada akhirnya membantu model klasifikasi agar tidak bias terhadap kelas dengan jumlah data lebih besar.



GAMBAR 4.3.
Grafik Akurasi Dalam Model LDA Setelah Oversampling

Laporan klasifikasi ini menunjukkan bahwa setelah menerapkan oversampling, model menunjukkan peningkatan kinerja dengan distribusi skor yang lebih seimbang di berbagai kelas. Precision tinggi pada beberapa kelas seperti 0 (41.67%), 3 (58.18%), dan 6 (75.00%), serta 7(64.91%), dan juga recall yang juga meningkat, terutama pada kelas-kelas yang sebelumnya bermasalah. Namun, masih ada kelas dengan kinerja rendah, seperti kelas 1, kelas 2, di mana precision, recall, dan F1-score tetap berada di bawah rata-rata. Oversampling membantu mengatasi ketidakseimbangan data antar kelas, sehingga model lebih

mampu mengenali pola pada kelas minoritas seperti 8 dan 10, meskipun kinerjanya masih belum optimal. Rata-rata makro kini menunjukkan precision, recall, dan F1-score yang lebih baik dibandingkan sebelum oversampling, mencerminkan peningkatan kemampuan model dalam menangani data yang lebih seimbang.



GAMBAR 4.4.
Grafik Hasil Confusion Matrix Setelah Oversampling

Confusion matrix menunjukkan bahwa model memiliki performa yang sangat tidak seimbang mesikpun telah memakai Oversampling, dengan bias yang signifikan terhadap kelas tertentu, terutama kelas 4, yang memiliki jumlah prediksi benar (true positives) paling tinggi dibandingkan kelas lainnya. Sebagian besar kelas lain, seperti kelas 8 dan 10, memiliki jumlah prediksi benar yang sangat kecil, mengindikasikan model sering salah memprediksi sampel dari kelas minoritas sebagai kelas mayoritas. Tingginya jumlah kesalahan klasifikasi terlihat dari banyaknya false positives dan false negatives di hampir semua kelas, menunjukkan bahwa model kesulitan mendiskriminasi antar kelas dengan baik. Ketidakseimbangan dalam distribusi data atau tumpang tindih fitur antar kelas kemungkinan besar menjadi penyebab utama model salah memetakan banyak prediksi ke kelas mayoritas, seperti kelas 4, sementara gagal mengenali pola pada kelas minoritas

TABEL 4.3.
Hasil Persentase Akurasi Keseluruhan Setelah Oversampling

	Precision	Recall	F1 – Score	Support
0	41.67%	14.71 %	21.74 %	102
1	29.27%	9.30%	14.12 %	129
2	25.64%	8.62%	12.90 %	116
3	58.18%	21.62 %	31.53 %	148
4	19.12%	61.70 %	29.19 %	141
5	20.83%	47.62 %	28.99 %	21
6	75.00%	45.00 %	56.25 %	20
7	64.91%	50.00 %	56.49 %	74
8	2.86%	12.50 %	4.65%	16
9	56.41%	35.48 %	43.56 %	62
10	0.00%	0.00%	0.00%	29
Accuracy	27.51%	27.51 %	27.51 %	858
Macro avg	35.81%	27.87 %	27.22 %	858
Weighted avg	37.98%	27.51 %	26.81 %	858

Hasil analisis data testing menggunakan oversampling :

a) Model menunjukkan bias yang signifikan terhadap kelas tertentu, khususnya kelas 4, yang terlihat dari tingginya Recall (61.70%) tetapi rendahnya Precision (19.12%).

- Interpretasi: Hal ini mengindikasikan bahwa model sering memprediksi kelas lain menjadi kelas 4. Meskipun model berhasil mengenali sebagian besar contoh kelas 4, banyak prediksi ke kelas 4 yang sebenarnya salah (false positives). Bias ini bisa terjadi akibat ketidakseimbangan distribusi data asli, bahkan setelah proses oversampling, atau karena kelas ini memiliki pola yang lebih mudah dikenali. Sebaliknya, kinerja model pada kelas minoritas, seperti kelas 8 dan kelas 10, sangat buruk. Hal ini terlihat dari Precision, Recall, dan F1-Score untuk kelas ini yang mendekati nol.
- Interpretasi: Ketidakmampuan model menangani kelas ini menunjukkan bahwa meskipun oversampling telah diterapkan, model tetap gagal mempelajari pola representatif untuk kelas minoritas. Kemungkinan penyebabnya adalah tumpang tindih fitur antara kelas minoritas dengan kelas mayoritas, sehingga model sulit membuat keputusan yang tepat.

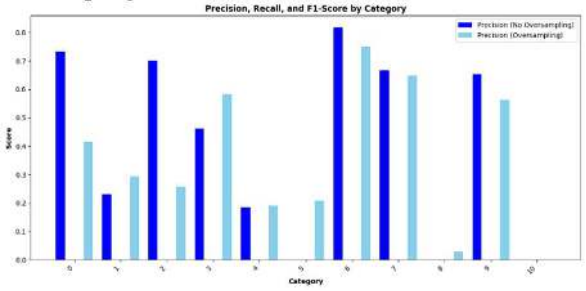
b) Kinerja model sangat dipengaruhi oleh overlap antar kelas. Model Linear Discriminant Analysis (LDA) menggunakan asumsi distribusi linier antar kelas, tetapi jika distribusi data antar kelas tidak linier atau saling tumpang tindih, model akan kesulitan mendiskriminasi kelas dengan baik. Kelas 6 memiliki kinerja terbaik dengan Precision 75.00% dan F1-Score 56.25%. Ini menunjukkan bahwa fitur kelas 6 mungkin memiliki distribusi yang lebih terpisah dibandingkan kelas lainnya. Sebaliknya, kelas 2, 8, dan 10 memiliki kinerja yang sangat rendah, yang menunjukkan overlap fitur yang parah atau fitur yang kurang representatif untuk kelas-kelas ini. Bias tinggi terhadap kelas tertentu (seperti kelas 4) juga memperkuat indikasi bahwa distribusi data atau fitur kelas mayoritas sangat mendominasi, sehingga kelas minoritas menjadi sulit dipelajar

Hasil evaluasi menunjukkan bahwa model LDA dengan oversampling memiliki kinerja yang rendah secara keseluruhan, dengan akurasi hanya 27.51%. Bias terhadap kelas mayoritas (kelas 4) menjadi tantangan utama, sementara kinerja pada kelas minoritas (kelas 8 dan 10) sangat buruk. Penyebab utama kegagalan ini adalah ketidakseimbangan data, distribusi fitur yang tumpang tindih, serta keterbatasan model LDA dalam menangkap pola non-linier walaupun sudah menggunakan Borderline-SMOTE untuk mengatasi ketidakseimbangan datanya.

4.3 Perbandingan Akurasi Oversampling dan Tidak Oversampling

1. Perbandingan dari segi F1-Score

Berikut adalah bar chart yang membandingkan precision antara model dengan oversampling dan tanpa oversampling berdasarkan kategori. Warna biru gelap merepresentasikan precision tanpa oversampling, sementara biru muda menunjukkan precision dengan oversampling.



.

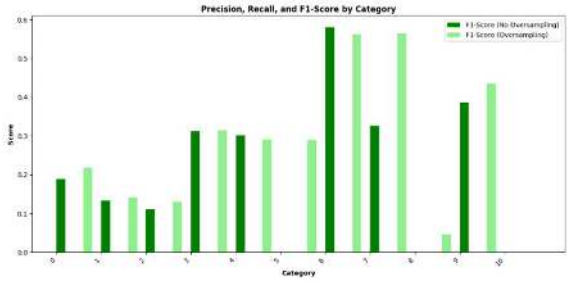
GAMBAR 4.5.
Grafik Oversampling vs No Oversampling model LDA by precision

Dari grafik ini, terlihat bahwa pada beberapa kategori, seperti kategori 0, 2, 5, dan 9, precision tanpa oversampling lebih tinggi dibandingkan dengan oversampling. Sebaliknya, pada kategori lain seperti 3, 4, 6, dan 8, oversampling justru meningkatkan precision. Dampak oversampling terhadap model cukup beragam. Secara umum, oversampling membantu meningkatkan precision pada kelas minoritas, yang sebelumnya kurang terwakili dalam data, namun pada beberapa kategori lainnya justru menurunkan precision. Hal ini mungkin terjadi karena oversampling membuat model lebih seimbang dalam mengenali berbagai kategori, tetapi juga berisiko menyebabkan overfitting pada beberapa kelas tertentu.

Dari hasil ini, oversampling memiliki dampak yang berbeda pada setiap kategori. Meskipun dapat membantu meningkatkan keseimbangan prediksi, efeknya terhadap keseluruhan kinerja model bervariasi. Oleh karena itu, penggunaan oversampling perlu dipertimbangkan dengan baik berdasarkan distribusi data dan tujuan model yang ingin dicapai

2. Perbandingan dari segi F1-Score

Berikut adalah bar chart yang membandingkan Precision, Recall, dan F1-Score berdasarkan kategori, terdapat dua kelompok bar yang merepresentasikan hasil tanpa oversampling (warna hijau tua) dan dengan oversampling (warna hijau terang).



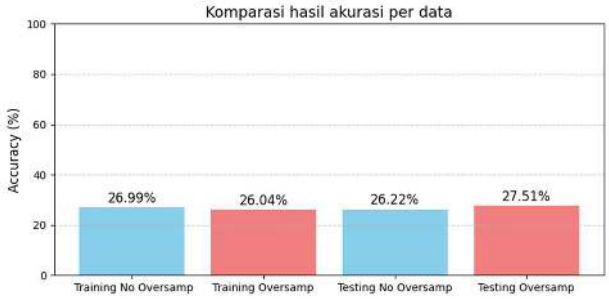
GAMBAR 4.6.
Grafik Oversampling vs No Oversampling model LDA by F1-Score

F1-Score dengan oversampling mengalami peningkatan pada banyak kategori dibandingkan dengan model tanpa oversampling. Hal ini menunjukkan bahwa oversampling membantu model dalam meningkatkan keseimbangan dalam memprediksi kategori yang sebelumnya kurang terwakili. Beberapa kategori mengalami kenaikan signifikan dalam F1-Score setelah dilakukan oversampling, terutama kategori 6, 7, dan 9. Hal ini menandakan bahwa model mendapatkan manfaat dari peningkatan representasi data dalam kategori-kategori tersebut.

Namun, ada kategori tertentu yang tidak menunjukkan perbedaan signifikan, seperti kategori 3 dan 4, di mana F1-Score tetap hampir sama antara kedua metode. Selain itu, kategori dengan skor rendah cenderung mengalami peningkatan setelah oversampling. Hal ini membuktikan bahwa metode ini efektif dalam mengatasi ketidakseimbangan data dan membantu model menghasilkan prediksi yang lebih baik secara keseluruhan. Oversampling membantu meningkatkan F1-Score pada sebagian besar kategori, terutama pada kelas yang sebelumnya memiliki jumlah sampel lebih sedikit. Peningkatan F1-Score ini menunjukkan bahwa model lebih baik dalam menangani ketidakseimbangan kelas setelah oversampling

3. METODE

Berikut adalah grafik menampilkan perbandingan akurasi antara model dengan dan tanpa oversampling dalam data training dan testing.



GAMBAR 4.7.
Grafik Perbandingan Akurasi Model LDA Oversampling Dan No oversampling

Akurasi pada data training tanpa oversampling sebesar 26.99%, sedangkan dengan oversampling sedikit turun menjadi 26.04%. Penurunan ini dapat terjadi karena

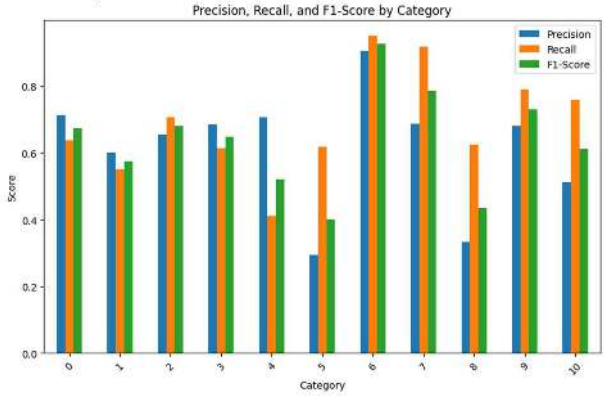
oversampling menambah data sintetik yang membuat model lebih general, tetapi tidak terlalu spesifik terhadap data training.

Sementara itu, akurasi pada data testing meningkat setelah oversampling, dari 26.22% menjadi 27.51%. Ini menunjukkan bahwa oversampling membantu meningkatkan performa model dalam menangani data yang tidak terlihat sebelumnya (testing), meskipun peningkatannya tidak terlalu besar. Peningkatan akurasi pada data testing lebih penting dibandingkan training karena menunjukkan bahwa model lebih mampu menangani data baru dengan lebih baik. Meskipun akurasi training sedikit menurun setelah oversampling, ini merupakan indikasi bahwa model menjadi lebih general dan tidak terlalu overfitting pada data latih.

4.4 Hasil pengujian metode pembanding XGBoost

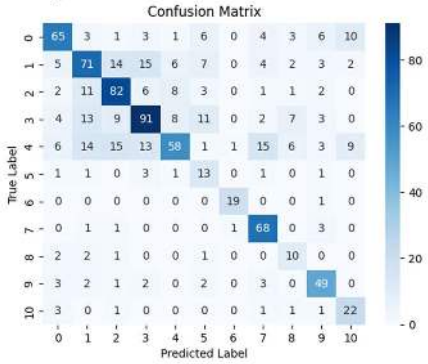
Pada bagian ini akan ditunjukkan grafik Precision, Recall, dan F1-Score untuk masing-masing kelas dalam hasil klasifikasi menggunakan XGBoost. Precision mengukur sejauh mana model menghindari prediksi positif yang salah, recall menunjukkan kemampuan model dalam menangkap semua sampel yang benar, sedangkan F1-Score merupakan keseimbangan antara precision dan recall. Rentang nilai pada grafik ini berkisar antara 0 hingga 1 yang menunjukkan nilai persentasenya, di mana skor yang lebih tinggi menunjukkan performa klasifikasi yang lebih baik.

Berikut adalah grafik yang menampilkan klasifikasi report untuk model XGBoost yang sudah di oversampling sebelumnya :



GAMBAR 4.8.
Grafik Akurasi Dalam Model XGBoost

Dari grafik, terlihat bahwa beberapa kelas memiliki skor yang tinggi dan stabil, sementara beberapa lainnya menunjukkan kinerja yang kurang optimal. Kelas 6 dan 7 memiliki Precision, Recall, dan F1-Score tertinggi, yang menunjukkan bahwa model dapat mengklasifikasikan kelas ini dengan sangat baik. Sebaliknya, kelas 4 dan 5 memiliki skor rendah di semua metrik, terutama F1-Score, yang menunjukkan bahwa model mengalami kesulitan dalam mengidentifikasi kelas ini secara akurat. Terdapat pula kelas dengan precision tinggi tetapi recall rendah, seperti kelas 3 dan 7, yang menunjukkan bahwa model mampu mengidentifikasi sampel yang benar dengan baik, tetapi masih banyak sampel dari kelas tersebut yang tidak terdeteksi. Sebaliknya, kelas 5 dan 9 memiliki recall tinggi tetapi precision rendah, yang berarti model sering menangkap sampel dari kelas ini tetapi juga menghasilkan banyak false positives.



GAMBAR 4.9.
Grafik Hasil Confusion Matrix Model XGBoost

Confusion matrix di atas menunjukkan kinerja model XGBoost dalam mengklasifikasikan berbagai kelas setelah dilakukan oversampling. Setiap sel pada matriks ini mewakili jumlah prediksi yang dibuat oleh model untuk setiap kombinasi antara label sebenarnya (True Label) dan label prediksi (Predicted Label). Diagonal utama dari matriks, yang berwarna lebih gelap, menunjukkan jumlah prediksi yang benar (True Positives) untuk setiap kelas. Semakin gelap warna pada diagonal, semakin tinggi jumlah prediksi yang benar, menandakan kinerja yang lebih baik dalam mengenali kelas tersebut. Cara membaca confusion matrix ini adalah dengan melihat nilai pada setiap sel. Misalnya, pada baris kelas 0, terdapat 65 perusahaan yang terprediksi benar (nilai diagonal) dan beberapa prediksi salah seperti 3 kali diklasifikasikan sebagai kelas 1 atau 6 kali sebagai kelas 4. Nilai-nilai ini membantu mengidentifikasi pola kesalahan spesifik, seperti apakah model cenderung mengklasifikasikan sampel ke kelas yang lebih dominan atau kelas dengan karakteristik yang mirip.

Dari confusion matrix ini, terlihat bahwa kelas 3, 6, dan 7 memiliki prediksi yang benar paling banyak, yaitu 91, 19, dan 68 sampel perusahaan yang diklasifikasikan dengan benar. Ini menunjukkan bahwa model XGBoost mampu mengenali pola dalam data kelas-kelas tersebut dengan baik, meskipun telah dilakukan oversampling. Namun, beberapa kelas seperti kelas 4 dan 5 mengalami banyak kesalahan klasifikasi, terlihat dari jumlah nilai tinggi di luar diagonal utama. Kelas 4, misalnya, sering salah diprediksi sebagai kelas 1, 2, dan 3, yang menunjukkan bahwa model kesulitan membedakan kelas ini dari kelas yang memiliki karakteristik serupa. Hal ini mengindikasikan bahwa meskipun oversampling telah dilakukan untuk menyeimbangkan data, model masih mengalami kesulitan dalam mengidentifikasi kelas minoritas dengan tepat.

TABEL 4. 4
Hasil Presentasi Akurasi Keseluruhan Model XGBoost

	Precision	Recall	F1 – Score	Support
0	71.43%	63.73 %	67.36 %	102

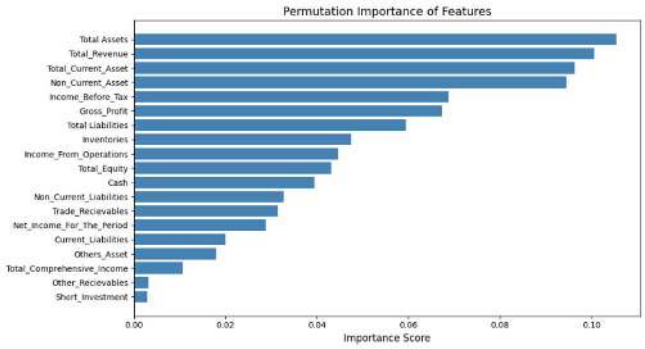
1	60.17%	55.04 %	57.49 %	129
2	65.60%	70.69 %	68.05 %	116
3	68.42%	61.49 %	64.77 %	148
4	70.73%	41.13 %	52.02 %	141
5	29.55%	61.90 %	40.00 %	21
6	90.48%	95.00 %	92.68 %	20
7	68.69%	91.89 %	78.61 %	74
8	33.33%	62.50 %	43.48 %	16
9	68.06%	79.03 %	73.13 %	62
10	51.16%	75.86 %	61.11 %	29
Accuracy	63.87%	63.87 %	63.87 %	858
Macro avg	61.60%	68.93 %	63.52 %	858
Weighted avg	65.86%	63.87 %	63.65 %	858

Berdasarkan tabel hasil evaluasi tersebut, model XGBoost menunjukkan performa yang lebih unggul dibandingkan dengan Linear Discriminant Analysis (LDA) dalam berbagai aspek evaluasi klasifikasi. Keunggulan utama XGBoost terlihat dari F1-score makro (Macro avg) sebesar 63,52%, yang menunjukkan bahwa model ini mampu menangani ketidakseimbangan kelas dengan lebih baik dibandingkan LDA, yang cenderung memiliki performa lebih rendah dalam situasi dengan distribusi kelas yang tidak merata. Selain itu, akurasi keseluruhan XGBoost mencapai 63,87%, yang biasanya lebih tinggi dibandingkan LDA dalam skenario klasifikasi non-linear
Dari segi precision dan recall, XGBoost menunjukkan keunggulan pada beberapa kelas tertentu, seperti kelas 6 dan 7, yang memiliki F1-score masing masing sebesar 92,68% dan 78,61%. Hal ini menunjukkan bahwa model mampu mengklasifikasikan beberapa kelas dengan sangat baik. LDA, yang bekerja berdasarkan asumsi distribusi normal antar kelas, sering kali kesulitan dalam menangani dataset dengan pola yang lebih kompleks, sedangkan XGBoost, sebagai model berbasis pohon keputusan yang diperkuat, lebih fleksibel dalam menangkap hubungan non-linear dalam data.

4.5 Hasil feature yang berpengaruh bagi kinerja model

1. Hasil feature yang berpengaruh untuk LDA

Berikut untuk grafik hasil analisis feature yang paling berpengaruh untul metode LDA menggunakan metode permutation importance :



GAMBAR 4.10.
Grafik Hasil Urutan Fitur Yang Paling Berpengaruh Untuk Model LDA

Berdasarkan hasil analisis feature importance, berikut adalah penjelasan terkait fitur-fitur yang memiliki pengaruh terhadap keberhasilan prediksi model LDA sebagai berikut :

Fitur-Fitur yang Paling Berdampak

a) Total_Assets: Jumlah total aset menunjukkan kapasitas keseluruhan perusahaan untuk menghasilkan pendapatan. Ini menjadi indikator yang penting untuk menilai kekuatan finansial perusahaan dalam sektor industri

b) Total_Revenue: Total pendapatan menunjukkan seberapa baik perusahaan menghasilkan pendapatan dari operasinya. Ini merupakan indikator langsung dari kinerja pasar dan efisiensi operasional.

c) Total_Current_Asset: Fitur ini memiliki skor penting tertinggi. Ini menunjukkan bahwa aset lancar total, seperti kas, piutang, atau inventaris yang mudah dikonversi menjadi uang tunai dalam waktu dekat, sangat memengaruhi keberhasilan prediksi. Hal ini relevan karena aset lancar mencerminkan likuiditas perusahaan, yang sering menjadi indikator kinerja dalam sektor industri.

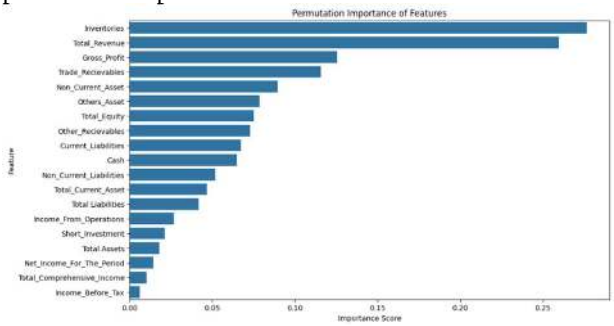
d) Non_Current_Asset: Aset tidak lancar, seperti properti, peralatan, atau aset jangka panjang lainnya, juga sangat signifikan. Ini karena aset ini mencerminkan investasi jangka panjang yang memengaruhi produktivitas dan kapasitas operasional perusahaan.

e) Income_Before_Tax: Pendapatan sebelum pajak mencerminkan profitabilitas perusahaan setelah dikurangi semua biaya operasional dan sebelum kewajiban pajak, yang menjadi faktor penting dalam menilai kesehatan keuangan.

Fitur-fitur dengan skor penting tinggi, terutama Total_Current_Asset, Non_Current_Asset, Total_Assets, dan Total_Revenue, harus menjadi fokus utama dalam analisis data keuangan untuk meningkatkan akurasi prediksi sektor industri. Fitur-fitur ini memberikan kontribusi signifikan karena mencerminkan elemen fundamental dari likuiditas, profitabilitas, dan investasi perusahaan, yang sangat berpengaruh dalam membedakan kinerja di berbagai sektor.

2. Hasil feature yang berpengaruh untuk XGBoost

Berikut untuk grafik hasil analisis feature yang paling berpengaruh untul metode LDA menggunakan metode permutation importance :



GAMBAR 4.11.
Grafik Hasil Urutan Fitur Yang Paling Berpengaruh Untuk Model XGBoost

Pada bagian ini, dilakukan juga analisis terhadap fitur yang paling berpengaruh dalam model XGBoost menggunakan metode Permutation Importance. Berdasarkan hasil pada gambar visualisasi Permutation Importance, fitur dengan importance score tertinggi untuk model XGBoost adalah sebagai berikut :

Fitur-Fitur yang Paling Berdampak

a) Inventories: Merupakan fitur dengan pengaruh terbesar terhadap model XGBoost. Fitur ini sangat relevan bagi perusahaan di sektor Consumer Goods, Manufaktur, dan Perdagangan, di mana tingkat persediaan barang menjadi faktor utama dalam operasional bisnis.

b) Total_Revenue: Pendapatan total mencerminkan performa keuangan perusahaan dan berpengaruh terhadap hampir semua sektor, terutama sektor Keuangan, Teknologi, dan Consumer Goods, yang memiliki pola pendapatan yang khas.

c) Gross_Profit: menunjukkan selisih antara pendapatan dan biaya pokok penjualan (COGS), yang mencerminkan profitabilitas operasional perusahaan sebelum memperhitungkan faktor lain seperti pajak dan beban operasional lainnya. Dalam model XGBoost, fitur ini berpengaruh signifikan karena perbedaan margin keuntungan antar sektor sangat jelas. Perusahaan dalam sektor Teknologi dan Healthcare cenderung memiliki gross profit tinggi, karena biaya produksi yang relatif lebih rendah dibandingkan pendapatannya, seperti pada perusahaan perangkat lunak dan farmasi.

d) Trade_Receivable: Fitur Trade Receivables (Piutang Dagang) dengan importance score 0.12 menggambarkan jumlah piutang dari pelanggan atas transaksi kredit, yang penting bagi sektor Keuangan, Properti, dan Perdagangan karena model bisnisnya yang mengandalkan pembayaran bertahap. Sebaliknya, sektor seperti Energi dan Consumer Goods memiliki piutang lebih rendah karena transaksi umumnya dilakukan secara tunai. Dengan perbedaan ini, XGBoost memanfaatkan fitur Trade Receivables untuk mengidentifikasi sektor yang lebih bergantung pada sistem kredit dibandingkan pembayaran langsung.

e) Non-Current_Asset: Fitur Non-Current Asset mencerminkan kepemilikan aset jangka panjang seperti tanah, bangunan, dan peralatan, yang dominan di sektor Infrastruktur, Properti, dan Energi, di mana aset fisik menjadi elemen utama bisnis. Sebaliknya, sektor Teknologi dan Jasa Keuangan lebih mengandalkan aset digital dan finansial. XGBoost menggunakan fitur ini untuk membedakan industri berbasis aset besar dari industri yang lebih bergantung pada modal intelektual atau digital.

Sementara itu, terdapat beberapa fitur dengan importance score yang lebih rendah (<0.02), seperti Income Before Tax dan Total Comprehensive Income, memiliki pengaruh minimal dalam menentukan sektor industri.

Hasil ini menunjukkan bahwa faktor utama dalam klasifikasi sektor industri lebih dipengaruhi oleh aspek aset dan pendapatan perusahaan dibandingkan dengan laba bersih atau pajak. Oleh karena itu, fitur dengan importance rendah dapat dipertimbangkan untuk dioptimasi atau dieliminasi dalam pemodelan lebih lanjut guna meningkatkan efisiensi model.

4.6 Analisis Perbandingan LDA dan XGBoost

Dari hasil skenario pengujian ini dimana akan mengetahui hasil dari masing masing pengujian yang diberikan terhadap prediksi sector ini yaitu dengan menggunakan model algoritma LDA dan membandingkan dengan metode pembandingnya yaitu mode XGBoost. Seeblumnya kedua dari model pengujian tersebut sudah dilakukan oversampling terlebih dahulu, dari kedua hasil masing – masing pengujian tersebut akan diperlihatkan perbandingannya:

TABEL 4. 5
Hasil Skenario Pengujian

NO	Skenario Pengujian	Hasil yang diharapkan	Hasil Pengujian	Ket
		LDA Model		
1	Melakukan prediksi model LDA dengan data yang sudah di oversampling	Hasil yang diharapkan akurasi dapat diatas 50%	Hasil yang didapatkan hanya sekitar 27.51%	Belum Sesuai yang diharapkan
		XGBoost Model		
2.	Melakukan prediksi model XGBoost dengan data yang sudah di oversampling	Hasil yang diharapkan akurasi dapat diatas 50%	Hasil yang didapatkan sekitar 63.87%	Sudah Sesuai yang diharapkan

Berdasarkan hasil tabel pengujian, terdapat perbedaan signifikan antara metode Linear Discriminant Analysis (LDA) dan Extreme Gradient Boosting (XGBoost) dalam memprediksi sektor industri berdasarkan laporan keuangan. XGBoost secara konsisten menunjukkan performa yang lebih baik dibandingkan LDA, baik dari segi akurasi, precision, recall, maupun F1-score.

Metode LDA hanya mencapai akurasi 27,51% setelah dilakukan oversampling, yang menunjukkan bahwa model ini kurang mampu menangkap pola kompleks dalam data keuangan. LDA bekerja dengan asumsi bahwa data terdistribusi secara linier dan antar kelas memiliki variansi yang sama, sehingga model ini memiliki keterbatasan dalam menangani data yang tidak memenuhi asumsi tersebut. Selain itu, nilai precision dan recall yang rendah pada banyak kelas menunjukkan bahwa LDA cenderung tidak mampu membedakan sektor dengan baik, terutama pada kelas yang memiliki jumlah data sedikit. Sebaliknya, XGBoost mencapai akurasi 63,87%, menunjukkan bahwa metode ini lebih mampu mengenali pola dalam data dengan baik. Sebagai model berbasis decision tree ensemble, XGBoost tidak memiliki asumsi distribusi linier, sehingga lebih fleksibel dalam menangani hubungan kompleks antar fitur keuangan. Selain itu, XGBoost memiliki F1-score makro sebesar 63,52%, yang berarti model ini lebih seimbang dalam mengklasifikasikan berbagai sektor industri, bahkan setelah menangani ketidakseimbangan data melalui oversampling.

Hasil pengujian juga menunjukkan bahwa beberapa kelas, seperti kelas 6 (Financials) dan kelas 7 (Properties & Real Estate), memiliki nilai precision dan recall tertinggi dalam XGBoost, menunjukkan bahwa model ini lebih dapat diandalkan dalam mengklasifikasikan sektor tertentu. Meskipun XGBoost jauh lebih unggul dibandingkan LDA, model ini masih mengalami beberapa tantangan, seperti kesalahan klasifikasi pada kelas minoritas.

5. KESIMPULAN

Berdasarkan hasil penelitian dan analisis yang dilakukan pada tugas akhir ini, implementasi Linear Discriminant Analysis (LDA) berhasil diterapkan sebagai metode klasifikasi sektor industri berdasarkan data laporan keuangan perusahaan yang terdaftar di Bursa Efek Indonesia (BEI). Model ini memanfaatkan fitur-fitur laporan keuangan sebagai prediktor, namun menunjukkan performa yang kurang optimal akibat tantangan distribusi data yang tidak seimbang. Hasil akurasi model sebesar 26,22% tanpa oversampling dan 27,51% setelah menggunakan oversampling mengindikasikan bahwa LDA kurang mampu menangkap pola klasifikasi yang kompleks. Hal ini menunjukkan bahwa meskipun oversampling telah meningkatkan distribusi data antar kelas, model tetap mengalami kesulitan dalam menangkap pola klasifikasi yang kompleks.

Sebagai pembanding, XGBoost menunjukkan akurasi lebih tinggi, yaitu 63,87%, menegaskan kemampuannya dalam menangani pola data yang lebih kompleks. Namun, penelitian ini tetap berfokus pada LDA, yang masih memiliki keterbatasan dalam klasifikasi data keuangan. Untuk penelitian selanjutnya, optimasi fitur dan eksplorasi metode lain dapat dilakukan guna meningkatkan performa klasifikasi.

Penelitian selanjutnya disarankan untuk menggunakan metode pembelajaran mesin yang lebih fleksibel, seperti Random Forest, Gradient Boosting, atau Neural Networks, yang mampu menangani pola non-linier dalam data dan memberikan akurasi prediksi yang lebih tinggi. Selain itu, peningkatan kualitas dataset dengan cakupan yang lebih luas, baik dari segi jumlah maupun periode waktu, serta penambahan fitur relevan lainnya juga dapat membantu meningkatkan akurasi prediksi. Untuk memastikan generalisasi model terhadap data yang tidak terlihat, teknik validasi seperti k-fold cross-validation atau stratified sampling dapat diterapkan. Dengan menerapkan saran-saran tersebut, diharapkan penelitian mendatang dapat menghasilkan model prediksi yang lebih akurat dan memberikan kontribusi yang lebih besar dalam pengelompokan sektor industri di Bursa Efek Indonesia.

Daftar Pustaka

[1] H. van der Heijden, “Predicting industry sectors from financial statements: An illustration of machine learning in accounting research,” British Accounting Review, vol. 54, no. 5, p. 101096, 2022, doi: 10.1016/j.bar.2022.101096.

[2] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

[3] Li, X., Liu, B., & Ding, Y. (2021). Bankruptcy Prediction Using XGBoost and Financial Ratios: A Comparative Study. Financial Data Science Journal, 12(2), 130-145.

[4] Anggraeni, R. D., Rahayu, S. M., & Topowijono. (2014). Penerapan Model Multiple Discriminant Analysis untuk Memprediksi Financial Distress (Studi pada Sektor Industri Barang Konsumsi yang Listing di Bursa Efek Indonesia Periode 2009-2012). Jurnal Administrasi Bisnis (JAB), 8(2). Retrieved from https://administrasibisnis.studentjournal.ub.ac.id

[5] Boedeker, P., & Kearns, N. T. (2019). Linear Discriminant Analysis for Prediction of Group Membership: A User-Friendly Primer. Advances in Methods and Practices in Psychological Science, 2(3), 250–263. https://doi.org/10.1177/2515245919849378

[6] Noviyanti Santoso dan Wahyu Wibowo, “Financial Distress Prediction using Linear Discriminant Analysis and Support Vector Machine”, 2018, Journal of Physics: Conference Series, Volume 979, The 2nd International Conference on Science (ICOS) 2–3 November 2017, Makassar, Indonesia.

[7] Nur Febrianti Bakri. (2024). Analisis Klasifikasi Financial Distress dengan Menggunakan Metode XGBoost. Universitas Hasanuddin.

[8] R. Dolphin, B. Smyth, and R. Dong, “A Machine Learning Approach to Industry Classification in Financial Markets,” in Artificial Intelligence and Cognitive Science, L. Longo and R. O’Reilly, Eds., Cham: Springer Nature Switzerland, 2023, pp. 81–94.

[9] Gischa, Serafica. Gischa, Serafica, ed. "Pengertian Laporan Keuangan, Tujuan dan Jenisnya". Kompas.com.

[10] Sunendar, Joeliardi (2019). Tim Sahamku, ed. Cara Mudah Memahami Laporan Keuangan. Joeliardi Sunendar. hlm. 17. ISBN 978-623-7231-18-9.

[11] Hidayat, Wastam Wahyu (2018). Fabri, Fungky, ed. Dasar-Dasar Analisa Laporan Keuangan . Ponorogo:

Uwais Inspirasi Indonesia. hlm. 3. ISBN 978 602-5891-76-2.
[12] Werner R. Murhadi, "Analisis Laporan Keuangan Proyeksi dan Valuasi Saham", 2019, Jakarta: Salemba Empat, ISBN 978-979-061-331-7
[13] Dian Ami Widyati, R. Rizal Isnanto, Munawar Agus Riyadi, " Analysis of Recognition Pattern Leaves uses the Method Linear Discriminant Analysis (LDA) and the Distance Minkowski ",TRANSFORMTIKA, Vol.18, No.2, January 2021, pp.225 – 230.
[14] Sebastian Raschka, and Vahid Mirjalili, " Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-learn, and TensorFlow2 ", 2019, Pack Publishing Ltd, Chapter 5 : Compressing Data via Dimensionality Reduction, Pages 159 -168
[15] Alaa Tharwat, Tarek Gaber, A. Ibrahim, and A. E.Hassanien, "Linear discriminant analysis: A detailed tutorial", 2017, AI Communications 30(2):169 190, DOI:10.3233/AIC-170729 [16] H. Yu and J. Yang, "A direct LDA algorithm for highdimensional data with application to face recognition", 2001, Pattern Recognition 34(10) (2001), 2067 2070. doi:10.1016/S0031- 3203(00)00162-X
[17] Christopher M. Bishop, " Pattern Recognition and Machine Learning" , Microsoft Research Ltd Cambridge CB3 0FB, U.K, 2006, (pp. 67-78).
[18] Shireen Elhabian and Aly A. Farag, "A Tutorial on Data Reduction: Linear Discriminant Analysis (LDA)", University of Louisville, CVIP Lab, 2009.
[19] Saptarsi Goswami, " Class Imbalance: SMOTE, Borderline-SMOTE, ADASYN ", Toward Data Science, 2019, Diakses pada 15 Januari 2025, dari https://towardsdatascience.com/class-imbalance-smote-borderline-smote adasyn-6e36c78d804.
[20] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. [DOI: 10.1613/jair.953]
[21] Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Advances in Intelligent Computing. Lecture Notes in Computer Science, vol. 3644, Springer. https://doi.org/10.1007/11538059_91
[22] Breiman, L. (2001). "Random Forests." Machine Learning, 45(1), 5-32. DOI: 10.1023/A:1010933404324.
[23] Reyvan Maulid, " Tools Data Science dengan Algoritma XGBoost ", 2023, Diakses pada 3 Februari 2025, https://dqlab.id/tools-data-science dengan-algoritma-xgboost?
[24] Tianqi Chen & Carlos Guestrin, " XGBoost: A Scalable Tree Boosting System ", University of Washington, (2016)
[25] Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.
[26] Lutifa Afifah, "Apa itu Confusion Matrix di Machine Learning?", Diakses pada 16 januari 2025, https://ilmudatapy.com/apa-itu-confusion-matrix