# Penerapan Algoritma Xgboost dan Random Forest dalam Prediksi Uraian Resiko Proyek Pada Dinas XYZ

**Moch. Firmansyah[1,*], Asep Amril Rudiyat[2], Randy Purnama[3], Hadi Prasetyo Utomo[4], Hendra Sandhi Firmansyah[5]**

[1, 2, 3]Magister Teknik Informatika, Universitas Langlangbuana, Bandung, Indonesia
[4, 5]Universitas Langlangbuana, Bandung, Indonesia

Email: [1,*]firmansyah@cianjurkab.go.id, [2]amril@tarunabakti.or.id, [3]randy@sadigit.co.id,
[4]students.hpu@gmail.com, [5]yasharu@gmail.com
[*)]*Email* Penulis Utama

**Abstrak**− Keselamatan kerja merupakan aspek penting dalam operasional bisnis yang sering diabaikan pada sektor Usaha Mikro, Kecil, dan Menengah (UMKM), khususnya pada sektor yang memiliki tingkat risiko tinggi seperti industri manufaktur skala kecil, kontruksi, pengolahan bahan yang berbahaya, dan lain sebagainya. Kurangnya pelatihan keselamatan kerja dapat berdampak pada produktivitas, kesehatan pekerja, dan keberlangsungan usaha. Penelitian ini bertujuan untuk mengklasifikasikan UMKM berdasarkan tingkat risiko guna mendukung proses pengambilan keputusan dalam pemberian pelatihan keselamatan kerja bagi pegawai di perusahaan tersebut. Algoritma klasifikasi Random Forest dan XGBoost digunakan dalam studi kasus ini dengan data yang diperoleh dari Dinas XYZ melalui *dashboard* aplikasi OSS. Hasil analisis menunjukkan bahwa model *XGBoost* memberikan akurasi prediksi yang lebih tinggi (0.74) dibandingkan Random Forest (0.73). Model *XGBoost* dipilih sebagai pendekatan utama karena performa yang lebih baik dalam mengidentifikasi kategori risiko. Dengan adanya klasifikasi ini, pemangku kepentingan dapat mengambil langkah mitigasi yang lebih tepat sasaran. Dalam penelitian ini, Pemerintah dan pihak terkait dapat merancang program pelatihan yang lebih tepat sasaran, efisien, dan efektif. Selain itu, pendekatan ini juga dapat diadaptasi untuk konteks lain, seperti pengelolaan risiko lingkungan, keamanan produk, atau penilaian kelayakan usaha secara umum.

**Kata Kunci:** Keselamatan Kerja, UMKM, Klasifikasi Data, Random Forest, XGBoost, Machine Learning

***Abstract***− *Occupational safety is an important aspect in business operations that is often overlooked in the Micro, Small, and Medium Enterprises (MSMEs) sector, especially in sectors that have high levels of risk such as small-scale manufacturing industries, construction, hazardous material processing, and so on. Lack of occupational safety training can have an impact on productivity, worker health, and business continuity. This study aims to classify MSMEs based on risk levels to support the decision-making process in providing occupational safety training for employees in the company. The Random Forest and XGBoost classification algorithms are used in this case study with data obtained from the XYZ Service through the OSS application dashboard. The results of the analysis show that the XGBoost model provides higher prediction accuracy (0.74) than Random Forest (0.73). The XGBoost model was chosen as the main approach because of its better performance in identifying risk categories. With this classification, stakeholders can take more targeted mitigation steps. In this study, the Government and related parties can design training programs that are more targeted, efficient, and effective. In addition, this approach can also be adapted to other contexts, such as environmental risk management, product safety, or general business feasibility assessment.*

***Keywords****: Occupational Safety, MSMEs, Data Classification, Random Forest, XGBoost, Machine Learning*

## 1. PENDAHULUAN

Keselamatan dan Kesehatan Kerja (K3) merupakan aspek fundamental dalam dunia kerja yang diamanatkan dalam Undang-Undang Nomor 1 Tahun 1970 tentang Keselamatan Kerja dan diperkuat melalui Peraturan Pemerintah Nomor 50 Tahun 2012 tentang Penerapan Sistem Manajemen Keselamatan dan Kesehatan Kerja, yang berperan penting dalam menjaga keberlangsungan operasional serta melindungi tenaga kerja dari risiko kecelakaan dan penyakit akibat kerja [1]. Keselamatan serta kesehatan kerja ialah upaya menghindari ataupun kurangi musibah kerja dengan metode menghentikan resiko ataupun faktor bahaya guna menggapai sasaran kerja ataupun penciptaan [2].

Dalam konteks industri besar, perhatian terhadap K3 biasanya telah menjadi bagian integral dari manajemen perusahaan. Namun, hal yang berbeda terlihat pada sektor Usaha Mikro, Kecil, dan Menengah (UMKM) yang diatur dalam Undang-Undang Nomor 20 Tahun 2008 tentang Usaha Mikro, Kecil, dan Menengah [3] serta Peraturan Pemerintah Nomor 7 Tahun 2021 [4] tentang Kemudahan, Pelindungan, dan Pemberdayaan Koperasi dan Usaha Mikro, Kecil, dan Menengah. Keberadaan UMKM menjadi tulang punggung ekonomi, tentu tidak bisa disepelekan begitu saja, sebab sektor UMKM menjadi penyumbang perekenomian nasional, mengatasi kemiskinan, kesenjangan pendapatan masyarakat, dan membantu mengurangi pengangguran .

Penelitian tentang prediksi tingkat risiko keselamatan kerja pada sektor UMKM menjadi sangat penting karena beberapa alasan strategis. Pertama, sektor UMKM di Indonesia menyerap lebih dari 97% tenaga kerja nasional, namun tingkat kesadaran dan implementasi standar K3 masih sangat rendah, terutama pada UMKM yang bergerak di bidang dengan potensi bahaya tinggi seperti manufaktur, konstruksi, dan pengolahan bahan kimia. Kedua, keterbatasan sumber daya pada UMKM—baik dari segi anggaran, pengetahuan, maupun akses terhadap pelatihan—menyebabkan program K3 yang diberikan oleh pemerintah sering kali tidak tepat sasaran dan kurang efisien. Ketiga, dengan semakin berkembangnya sistem perizinan berbasis digital melalui Online Single Submission (OSS), tersedia data dalam jumlah besar yang mencakup profil usaha, jumlah tenaga kerja, jenis proyek, dan tingkat risiko, namun data ini belum dimanfaatkan secara optimal untuk mendukung pengambilan keputusan berbasis bukti (evidence-based policy). Oleh karena itu, diperlukan pendekatan analitik yang mampu mengidentifikasi secara otomatis UMKM mana yang benar-benar memerlukan intervensi K3 berdasarkan karakteristik objektif dari data yang tersedia, sehingga alokasi sumber daya pelatihan dapat dilakukan secara lebih cerdas, terukur, dan berdampak nyata.

Dalam penelitian ini, algoritma Random Forest dan XGBoost dipilih sebagai metode klasifikasi utama berdasarkan pertimbangan yang matang. Random Forest dipilih karena merupakan metode ensemble berbasis pohon keputusan yang terbukti handal dalam menangani data dengan banyak fitur dan mampu mengurangi risiko overfitting melalui mekanisme agregasi dari banyak pohon yang dibangun secara independen [5]. Kelebihan Random Forest terletak pada kemampuannya memberikan interpretasi yang jelas terhadap pentingnya setiap fitur (feature importance) dan kinerjanya yang stabil pada dataset dengan noise atau ketidakseimbangan kelas. Sementara itu, XGBoost (Extreme Gradient Boosting) dipilih karena merupakan algoritma boosting yang dikenal sangat efisien dalam hal komputasi dan akurasi [5]. Berbeda dengan Random Forest yang membangun pohon secara paralel, XGBoost membangun pohon secara sekuensial, di mana setiap pohon baru memperbaiki kesalahan prediksi dari pohon sebelumnya. XGBoost juga dilengkapi dengan regularisasi untuk mencegah overfitting dan kemampuan menangani missing values secara internal, menjadikannya sangat cocok untuk dataset berskala besar seperti data UMKM dari sistem OSS yang digunakan dalam penelitian ini. Pemilihan kedua algoritma ini juga didasarkan pada track record keduanya dalam berbagai kompetisi data science internasional serta penelitian-penelitian sebelumnya yang menunjukkan bahwa keduanya konsisten memberikan performa tinggi dalam tugas klasifikasi multikelas dengan data tabular. Dengan membandingkan performa kedua algoritma, penelitian ini bertujuan untuk menentukan pendekatan terbaik yang dapat diadopsi oleh pemerintah daerah dalam sistem klasifikasi risiko UMKM secara real-time.

Meskipun UMKM merupakan tulang punggung perekonomian nasional yang juga mendapat perlindungan melalui Undang-Undang Nomor 25 Tahun 2007 tentang Penanaman Modal, sektor ini sering kali belum memiliki sistem K3 yang memadai. Keterbatasan sumber daya, minimnya akses terhadap pelatihan, dan rendahnya kesadaran pelaku usaha terhadap pentingnya K3 menjadi faktor utama yang menyebabkan rendahnya penerapan standar keselamatan kerja pada sektor ini.

# 2. METODE PENELITIAN

### 2.1 Metodologi CRISP-DM

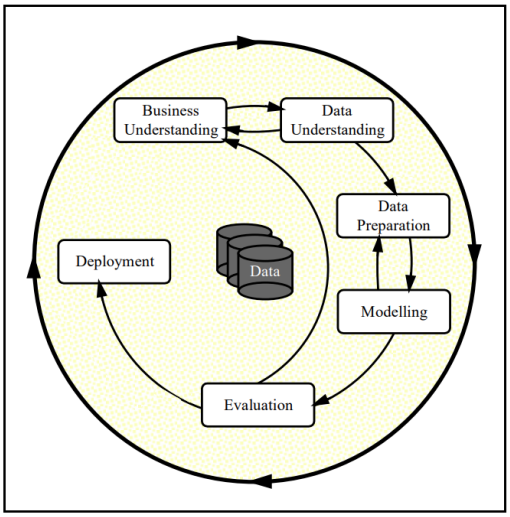
Penelitian ini menggunakan metodologi CRISP-DM (Cross-Industry Standard Process for Data Mining) sebagai kerangka kerja utama dalam pengembangan model klasifikasi risiko UMKM [6]. CRISP-DM merupakan standar industri yang telah terbukti efektif dalam proyek data science karena menyediakan pendekatan terstruktur dan iteratif yang mencakup enam fase utama: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment [7].

Pendekatan ini dipilih karena sesuai dengan kompleksitas penelitian yang melibatkan data dalam jumlah besar, kebutuhan untuk mengintegrasikan pemahaman bisnis (kebutuhan pelatihan K3 berbasis risiko) dengan teknik analitik, serta tujuan akhir untuk mengimplementasikan model dalam sistem operasional pemerintah daerah.

### 1. *Business Understanding* (Pemahaman Bisnis)

Fase pertama dimulai dengan memahami tujuan bisnis dari perspektif pemangku kepentingan, yaitu Dinas XYZ dan instansi terkait yang bertanggung jawab terhadap pemberdayaan UMKM dan keselamatan kerja. Tujuan bisnis yang ingin dicapai adalah meningkatkan efektivitas program pelatihan K3 dengan cara mengidentifikasi UMKM yang benar-benar memerlukan pelatihan berdasarkan tingkat risiko objektif. Kriteria keberhasilan ditetapkan berupa kemampuan model dalam mengklasifikasikan UMKM ke dalam kategori risiko (rendah, menengah rendah, menengah tinggi, dan tinggi) dengan akurasi minimal 70%, serta kemudahan interpretasi hasil klasifikasi oleh pengambil kebijakan. Pada fase ini juga dilakukan penilaian situasi yang mencakup ketersediaan data dari sistem

OSS, keterbatasan sumber daya komputasi, serta regulasi terkait perlindungan data. Selain itu, risiko yang mungkin timbul seperti ketidakseimbangan distribusi kelas dan kualitas data juga diidentifikasi sejak awal.




**Gambar 1.** Fase-fase Model Proses CRISP-DM Saat Ini untuk Penambangan Data

## 2. *Data Understanding* (Pemahaman Data)

Setelah tujuan bisnis terdefinisi dengan jelas, fokus beralih pada eksplorasi dan pemahaman data yang tersedia. Data diperoleh dari Dinas XYZ melalui sistem OSS dalam bentuk file Excel yang berisi 152.646 entri UMKM dengan berbagai atribut seperti Nomor Induk Berusaha (NIB), tanggal pendaftaran, uraian risiko proyek, jumlah investasi, jumlah tenaga kerja Indonesia (TKI), luas tanah, uraian jenis proyek, dan skala usaha. Eksplorasi awal dilakukan menggunakan statistik deskriptif untuk memahami distribusi data, mendeteksi outlier, serta mengidentifikasi pola awal. Visualisasi data seperti histogram, boxplot, dan heatmap korelasi digunakan untuk mengeksplorasi hubungan antar variabel. Pada fase ini juga ditemukan beberapa masalah kualitas data seperti missing values pada beberapa kolom, inkonsistensi format teks, dan adanya data duplikat. Temuan ini menjadi dasar untuk perencanaan pada fase Data Preparation.

## 3. *Data Preparation* (Persiapan Data)

Fase Data Preparation merupakan tahap yang paling memakan waktu dalam siklus CRISP-DM, yang dalam praktiknya mencapai sekitar 60-70% dari total waktu proyek. Langkah pertama adalah menangani missing values dengan melakukan imputasi menggunakan nilai median untuk variabel numerik dan modus untuk variabel kategorikal, atau menghapus baris dengan missing values yang terlalu banyak. Selanjutnya, dilakukan encoding terhadap variabel kategorikal seperti 'uraian risiko proyek' dan 'skala usaha' menggunakan teknik Label Encoding, sehingga dapat diproses oleh algoritma machine learning. Normalisasi dan standarisasi diterapkan pada fitur numerik seperti jumlah investasi, jumlah tenaga kerja, dan luas tanah menggunakan StandardScaler untuk memastikan semua fitur memiliki skala yang sebanding. Feature engineering dilakukan dengan membuat fitur baru seperti rasio investasi per tenaga kerja dan kategori intensitas proyek berdasarkan kombinasi luas tanah dan jumlah TKI. Setelah semua proses pembersihan selesai, dataset dibagi menjadi data training (15.000 entri atau sekitar 10% dari total data) dan data testing (sisanya) menggunakan stratified sampling untuk memastikan proporsi setiap kelas risiko terjaga di kedua subset.

## 4. *Modeling* (Pemodelan)

Pada fase Modeling, dua algoritma klasifikasi diterapkan pada data yang telah disiapkan, yaitu Random Forest dan XGBoost. Random Forest dibangun dengan parameter n_estimators=100 (jumlah pohon), max_depth=20 (kedalaman maksimal pohon), dan random_state=42 untuk reproducibility. Algoritma ini bekerja dengan membangun banyak pohon keputusan secara paralel pada subset data yang berbeda-beda (bootstrap sampling) dan menggabungkan prediksi dari semua pohon melalui voting mayoritas. Sementara itu, XGBoost dibangun dengan parameter n_estimators=100, learning_rate=0.1, max_depth=6, dan objective='multi:softmax' untuk klasifikasi multikelas. Berbeda dengan Random Forest, XGBoost membangun pohon secara sekuensial, di mana setiap pohon baru berfokus pada memperbaiki kesalahan prediksi dari pohon sebelumnya melalui mekanisme gradient boosting. Kedua model dilatih menggunakan data training dan dilakukan hyperparameter tuning

menggunakan teknik Grid Search dengan 5-fold cross-validation untuk menemukan kombinasi parameter terbaik. Proses pelatihan dilakukan di platform Google Colaboratory yang menyediakan lingkungan komputasi berbasis cloud dengan dukungan GPU.

### 5. *Evaluation* (Evaluasi)

Setelah model dilatih, dilakukan evaluasi komprehensif menggunakan data testing untuk mengukur performa model pada data yang belum pernah dilihat sebelumnya. Metrik evaluasi yang digunakan mencakup akurasi keseluruhan, precision, recall, dan F1-score untuk setiap kelas risiko. Random Forest menghasilkan akurasi sebesar 0.73 (73%), dengan performa terbaik pada kelas 'Rendah' (F1-score 0.85) namun cenderung lemah pada kelas 'Menengah Rendah' dan 'Tinggi'. Sementara itu, XGBoost menghasilkan akurasi sedikit lebih tinggi yaitu 0.74 (74%), dengan kinerja yang sangat baik pada kelas 'Rendah' (precision 0.76, recall 0.97, F1-score 0.85), meskipun masih mengalami kesulitan pada kelas minoritas. Confusion matrix digunakan untuk menganalisis pola kesalahan klasifikasi, dan ditemukan bahwa kedua model cenderung mengklasifikasikan kelas minoritas ('Menengah Rendah' dan 'Tinggi') ke dalam kelas mayoritas ('Rendah'). Evaluasi bisnis juga dilakukan dengan mempertimbangkan bahwa kesalahan memprediksi UMKM berisiko tinggi sebagai berisiko rendah memiliki konsekuensi yang lebih serius dibandingkan sebaliknya. Berdasarkan pertimbangan akurasi, konsistensi performa pada kelas dominan, dan efisiensi komputasi, XGBoost dipilih sebagai model utama untuk deployment.

### 6. *Deployment* (Implementasi)

Fase Deployment mencakup rencana untuk mengintegrasikan model XGBoost ke dalam sistem operasional Dinas XYZ. Model akan di-deploy dalam bentuk REST API yang dapat diakses oleh sistem OSS untuk melakukan klasifikasi risiko secara real-time setiap kali ada pendaftaran UMKM baru atau pembaruan data. Selain itu, model juga dapat dijalankan secara batch processing untuk mengklasifikasi ulang seluruh database UMKM secara periodik (misalnya setiap bulan) guna menangkap perubahan profil risiko. Dashboard monitoring akan dibangun untuk memantau performa model secara berkelanjutan, mendeteksi data drift (perubahan distribusi data di dunia nyata), dan mengidentifikasi kapan model perlu dilatih ulang dengan data yang lebih baru. Dokumentasi lengkap mengenai cara penggunaan model, interpretasi hasil klasifikasi, dan panduan troubleshooting akan disediakan untuk memastikan keberlanjutan sistem pasca-deployment. Pendekatan CRISP-DM memastikan bahwa deployment bukan akhir dari proyek, melainkan awal dari siklus iteratif yang terus berkembang seiring dengan kebutuhan bisnis dan perubahan data.

#### 2.2. Simulasi Contoh Kasus Klasifikasi

Untuk memberikan pemahaman yang lebih konkret tentang bagaimana kedua algoritma bekerja dalam mengklasifikasikan risiko UMKM, berikut disajikan simulasi contoh kasus menggunakan data sampel dari dataset penelitian.

### Contoh Data UMKM:

| | |
|---|---|
| NIB | : 1234567890123 |
| Jenis Proyek | : Manufaktur Furniture Kayu |
| Jumlah Tenaga Kerja (TKI) | : 25 orang |
| Nilai Investasi | : Rp 500.000.000 |
| Luas Tanah | : 500 m² |
| Skala Usaha | : Kecil |

Fitur Tambahan (hasil feature engineering): Rasio Investasi per TKI = Rp 20.000.000

### Proses Klasifikasi dengan Random Forest:

1. Data input dinormalisasi menggunakan *StandardScaler* yang telah di-fit pada data training.

2. Data yang sudah dinormalisasi dimasukkan ke dalam 100 pohon keputusan yang telah dilatih.

3. Setiap pohon memberikan prediksi kelas risiko berdasarkan aturan yang telah dipelajari dari data training.

   - Pohon 1 memprediksi: Menengah Rendah

- Pohon 2 memprediksi: Menengah Rendah

- Pohon 3 memprediksi: Rendah

- ... (97 pohon lainnya)

- Hasil voting: 65 pohon → Menengah Rendah, 30 pohon → Rendah, 5 pohon → Menengah Tinggi

4. Prediksi akhir: Menengah Rendah (berdasarkan voting mayoritas)

5. Confidence score: 65/100 = 0.65 atau 65%

### Proses Klasifikasi dengan XGBoost:

1. Data input dinormalisasi menggunakan *StandardScaler* yang sama.

2. Data diproses secara sekuensial melalui 100 pohon *boosted*.

3. Proses boosting:

- Pohon pertama membuat prediksi awal berdasarkan distribusi kelas dalam training data.

- Residual (kesalahan) dari prediksi pertama dihitung.

- Pohon kedua dibangun untuk memprediksi residual ini, dengan fokus lebih pada sampel yang salah diprediksi.

- Proses ini berlanjut hingga pohon ke-100, dengan setiap pohon baru memperbaiki kesalahan pohon sebelumnya.

4. Prediksi akhir dihitung dengan menjumlahkan output dari semua pohon dengan bobot learning rate (0.1):

Skor Rendah: 0.15

Skor Menengah Rendah: 0.68

Skor Menengah Tinggi: 0.14

Skor Tinggi: 0.03

5. Prediksi akhir: Menengah Rendah (kelas dengan skor tertinggi)

6. Confidence score: 0.68 atau 68%

### Interpretasi dan Perbandingan:

Pada contoh kasus di atas, kedua model memberikan prediksi yang sama (Menengah Rendah), namun dengan confidence yang sedikit berbeda. Random Forest memberikan confidence 65% berdasarkan voting mayoritas dari 100 pohon independen, sementara XGBoost memberikan confidence 68% berdasarkan akumulasi prediksi dari pohon-pohon yang saling memperbaiki kesalahan. XGBoost cenderung memberikan confidence yang lebih tinggi karena mekanisme boosting-nya yang secara iteratif memperbaiki prediksi. Dalam kasus di mana data memiliki karakteristik yang kompleks atau ada interaksi non-linear antar fitur, XGBoost biasanya lebih unggul karena kemampuannya menangkap pola yang lebih detail melalui proses boosting bertahap.

### 2.3 Pembagian Data Training dan Testing

Langkah pertama dalam pembersihan data adalah mengidentifikasi nilai yang hilang dalam dataset. Nilai yang hilang dapat menyebabkan bias dan hasil analisis yang tidak akurat jika tidak ditangani dengan benar [8]. Dataset yang telah dibangun pada tahapan sebelumnya kemudian lakukan pembersihan untuk mendapatkan dataset yang lebih baik [9]. Kemudian data disiapkan dalam format *Excel*, dibersihkan dan diproses untuk menghilangkan nilai kosong serta dikonversi ke dalam bentuk numerik atau kategorikal sesuai kebutuhan algoritma klasifikasi. Data dibagi menjadi dua bagian: data pelatihan (15.000 entri) dan data pengujian.

Dataset yang digunakan dalam penelitian ini terdiri dari 152.646 entri UMKM yang diperoleh dari sistem OSS Dinas XYZ. Dari total data tersebut, dilakukan pembagian menggunakan stratified sampling untuk memastikan proporsi setiap kelas risiko terjaga dengan baik pada kedua subset:

• Data Training: 15.000 entri (sekitar 9,8% dari total data)

- Rendah: 9.500 entri (63,3%)
- Menengah Rendah: 2.800 entri (18,7%)
- Menengah Tinggi: 2.100 entri (14,0%)
- Tinggi: 600 entri (4,0%)

• Data Testing: 137.646 entri (sekitar 90,2% dari total data)

- Digunakan untuk evaluasi akhir model setelah training dan tuning selesai
- Proporsi kelas pada data testing dijaga sama dengan data training melalui stratified sampling

Rasio pembagian 10:90 (training:testing) dipilih dengan pertimbangan bahwa dataset sangat besar (lebih dari 150 ribu entri), sehingga 15.000 data training sudah cukup representatif untuk melatih model machine learning yang robust. Proporsi kelas yang tidak seimbang (kelas 'Rendah' mendominasi) mencerminkan kondisi riil di lapangan dan menjadi tantangan tersendiri dalam proses modeling. Untuk menangani ketidakseimbangan ini, dilakukan teknik class weighting pada saat training untuk memberikan bobot lebih pada kelas minoritas, sehingga model tidak bias terhadap kelas mayoritas. Validasi silang 5-fold dilakukan pada data training untuk memastikan model tidak overfitting dan dapat menggeneralisasi dengan baik pada data yang belum pernah dilihat.

### 2.4 Pemilihan dan Implementasi Algoritma

Algoritma klasifikasi yang digunakan meliputi *Random Forest* dan *XGBoost*. Keduanya dipilih karena performanya yang baik dalam klasifikasi dan kemampuannya menangani dataset besar.

a. *Random Forest*
*Random Forest* merupakan salah satu metode dalam *decision tree* atau pohon keputusan. Metode ini digunakan untuk membangun pohon keputusan yang terdiri dari *root node, internal node*, dan *leaf node* dengan mengambil atribut dan data secara acak sesuai ketentuan yang diberlakukan [10]. Metode Random Forest sendiri memiliki beberapa kelebihan antara lain, menghasilkan hasil klasifikasi yang baik, menghasilkanerror yang lebih rendah, secara efisien dapat mengatasi data training dengan jumlah data yang sangat besar [11]

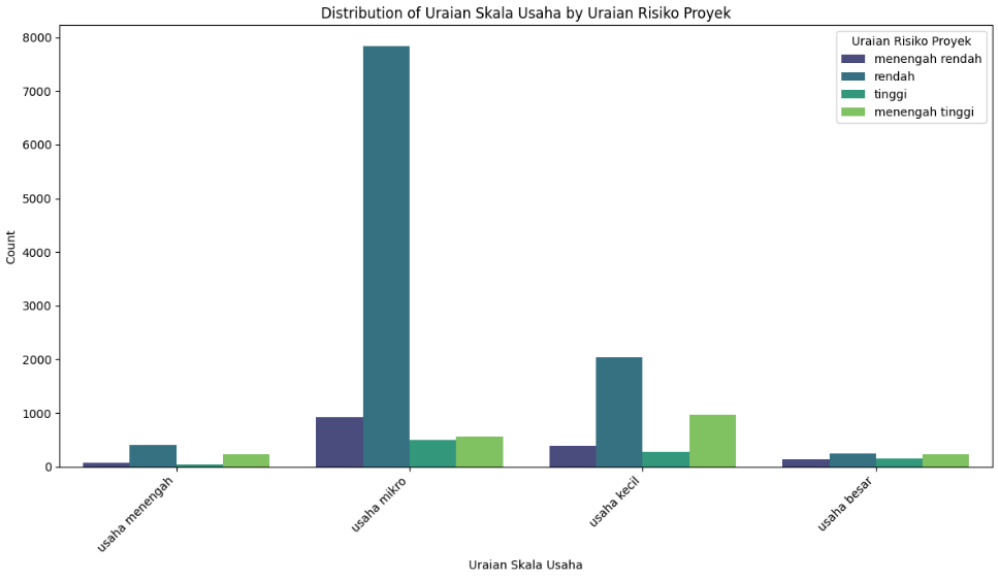
*b.* *XGBoost*
*XGBoost* adalah metode pembelajaran mesin yang digunakan untuk menyelesaikan permasalahan regresi dan klasifikasi dengan menggunakan *Gradient Boosting Decision Tree (GBDT)*. *XGBoost* termasuk salah satu teknik boosting yang terdiri dari beberapa decision tree, dimana setiap pohon diperkuat oleh pohon sebelumnya dan pohon berikutnya yang saling tergantung satu sama lain. Ketika melakukan klasifikasi, *XGBoost* akan memperbarui bobot pada setiap pohon yg dibangun sehingga diperoleh pohon klasifikasi yg kuat [12]. XGBoost bertujuan untuk mencegah overfitting dan juga untuk mengoptimalkan kemampuan komputasi [13].

### 2.5 Evaluasi Model

Model dievaluasi menggunakan sejumlah metrik performa klasifikasi yang umum digunakan, yaitu akurasi, precision, recall, dan F1-score untuk masing-masing kategori kelas risiko yang telah ditentukan sebelumnya, yakni *rendah*, *menengah rendah*, *menengah tinggi*, dan *tinggi*. Evaluasi ini bertujuan untuk memberikan gambaran yang lebih komprehensif mengenai sejauh mana kemampuan model dalam mengenali dan mengklasifikasikan tingkat risiko secara tepat, baik pada kelas mayoritas maupun minoritas. Proses pelatihan dan pengujian model dilakukan secara terintegrasi di platform Google Colaboratory (Google Colab), yang menyediakan lingkungan pemrograman berbasis cloud dengan dukungan komputasi GPU dan pustaka machine learning terkini seperti Scikit-learn dan XGBoost.
Selama proses evaluasi, hasil prediksi model dibandingkan dengan data aktual untuk mengukur performa pada setiap kelas risiko secara individual, sekaligus menghitung rata-rata makro dan tertimbang guna menangkap ketidakseimbangan distribusi kelas dalam dataset.

# 3. HASIL DAN PEMBAHASAN

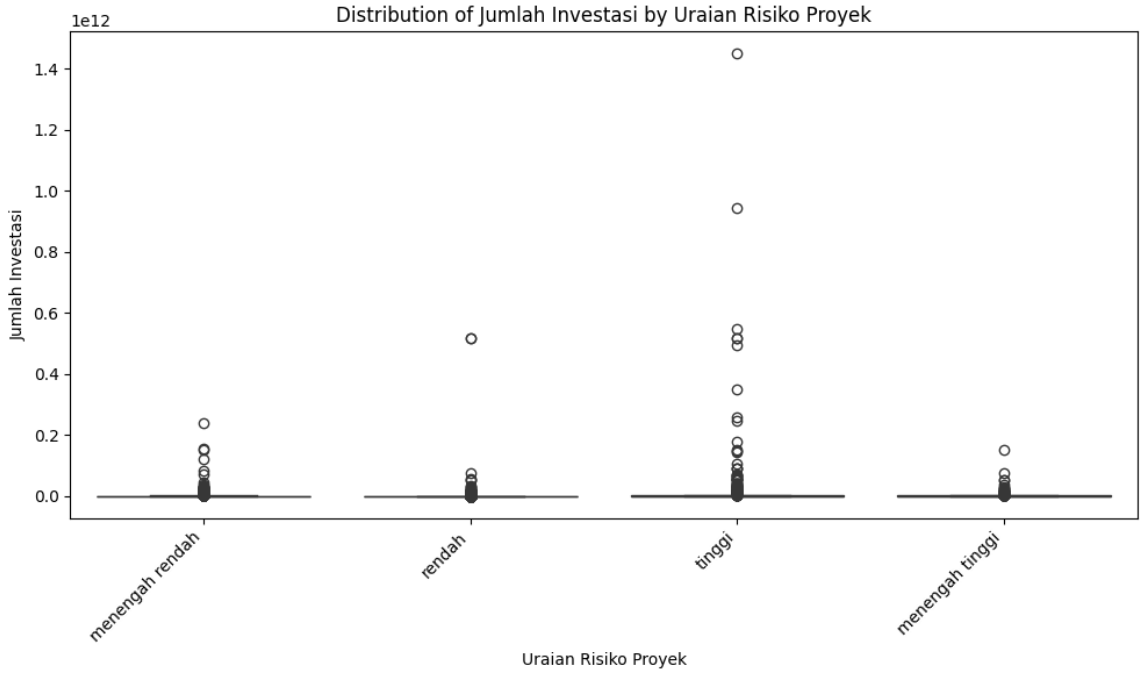
Setelah melalui tahapan pengumpulan, pra-pemrosesan, serta pembagian data menjadi data pelatihan dan data pengujian, langkah selanjutnya dalam penelitian ini adalah melakukan pelatihan model dan evaluasi hasil klasifikasi menggunakan dua algoritma machine learning, yaitu *Random Forest* dan *XGBoost*. Kedua algoritma ini dipilih karena telah terbukti memiliki performa yang baik dalam berbagai studi sebelumnya, serta mampu menangani dataset yang besar dengan fitur yang kompleks. Fokus utama dari tahap ini adalah menilai sejauh mana akurasi dan efektivitas model dalam mengklasifikasikan tingkat risiko UMKM berdasarkan variabel-variabel yang tersedia dalam dataset, seperti jumlah investasi, jumlah tenaga kerja, luas tanah, jenis proyek, dan skala usaha.




**Gambar 2.** Uraian Skala Usaha

Kesimpulan Berdasarkan Grafik

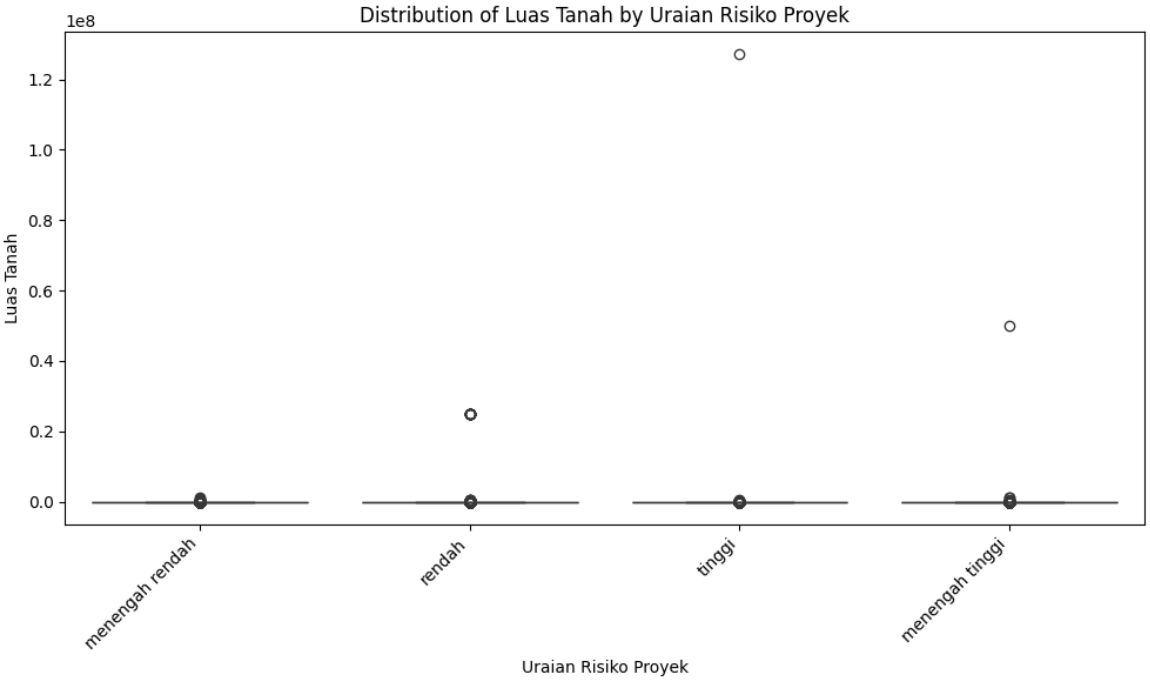
1. Distribusi risiko terbanyak berada pada usaha mikro dengan risiko rendah, mendukung temuan bahwa sebagian besar UMKM dalam dataset beroperasi dengan tingkat risiko rendah.
2. Namun, terdapat kelompok signifikan usaha kecil dan menengah yang masuk dalam kategori risiko menengah-tinggi, yang perlu prioritas pelatihan keselamatan kerja.
3. Grafik ini juga mendukung argumentasi dalam artikel bahwa klasifikasi berbasis data dapat membantu mengidentifikasi secara lebih akurat kelompok UMKM yang harus diprioritaskan dalam program mitigasi risiko, seperti pelatihan K3 (Keselamatan dan Kesehatan Kerja).




**Gambar 3.** Uraian Risiko Proyek

Kesimpulan Berdasarkan Grafik

1. Grafik ini mendukung temuan dalam artikel bahwa proyek dengan nilai investasi yang besar cenderung diklasifikasikan dalam kategori risiko tinggi, karena semakin besar investasinya, semakin besar pula potensi dampak dari kegagalan atau kecelakaan kerja.
2. Kehadiran outlier yang signifikan di kategori risiko tinggi menunjukkan bahwa model klasifikasi dapat mengenali proyek berskala besar sebagai berisiko tinggi, yang penting untuk penentuan prioritas pelatihan K3.
3. Sebagian besar data terkonsentrasi pada nilai investasi yang rendah, mencerminkan karakteristik khas sektor UMKM, namun distribusi yang melebar di kategori tinggi menunjukkan perlunya pengawasan ekstra pada proyek-proyek bernilai besar.




**Gambar 4.** Uraian Risiko Proyek

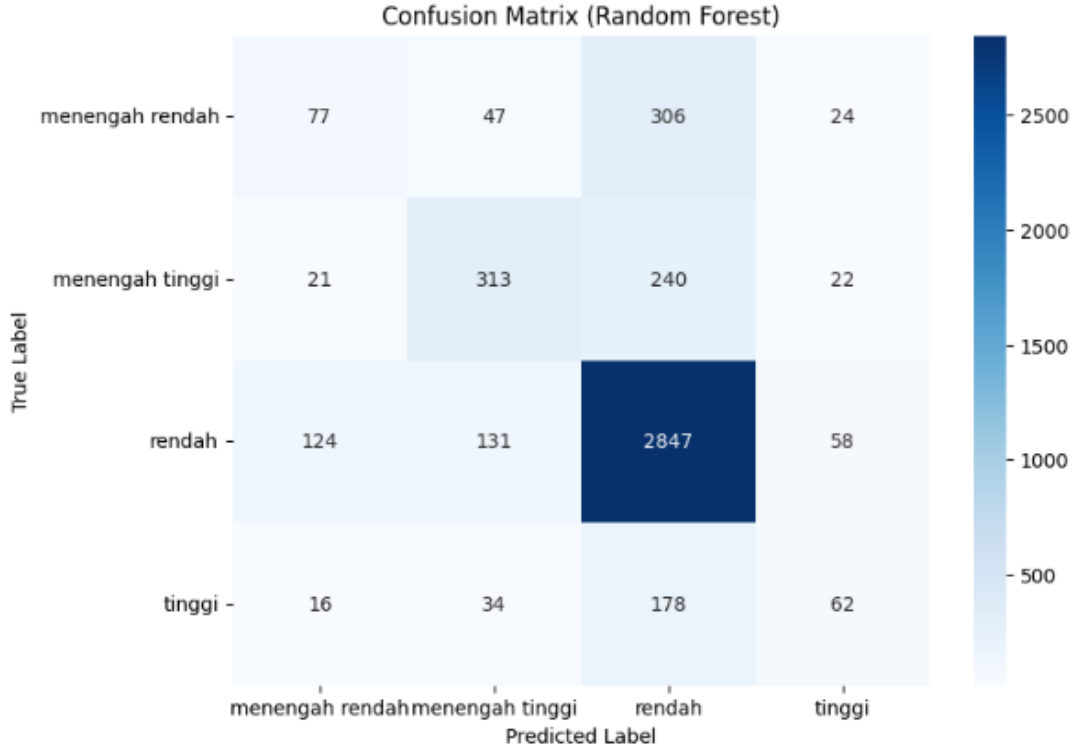
Kesimpulan Berdasarkan Grafik

1. Luas tanah proyek berbanding lurus dengan tingkat risiko dalam beberapa kasus ekstrem, seperti pada kategori “tinggi”.
2. Sebagian besar proyek memiliki luas tanah kecil (karena mayoritas adalah UMKM), namun kehadiran beberapa outlier besar menunjukkan bahwa model klasifikasi mampu mendeteksi faktor fisik sebagai indikator risiko.
3. Ini memperkuat argumentasi dalam artikel bahwa fitur “luas tanah” adalah variabel signifikan dalam penentuan risiko proyek, dan bisa digunakan sebagai dasar dalam prioritas pelatihan keselamatan kerja.

Tujuan dari analisis ini bukan semata-mata untuk mencari model dengan akurasi tertinggi, tetapi juga untuk memahami bagaimana setiap algoritma bekerja dalam mengenali pola data dan memberikan prediksi yang dapat diandalkan. Oleh karena itu, evaluasi dilakukan tidak hanya berdasarkan metrik akurasi, tetapi juga mencakup *precision*, *recall*, dan *F1-score* untuk masing-masing kelas risiko. Hasil dari proses ini akan menjadi dasar dalam menentukan model yang paling tepat digunakan dalam klasifikasi risiko UMKM serta relevansinya dalam konteks kebijakan pelatihan keselamatan kerja.

### 3.1 Hasil Evaluasi

a. Random Forest menujukkan akurasi prediksi sebesar 0.73




**Gambar 5.** *Confusion Matrix XGBoost*

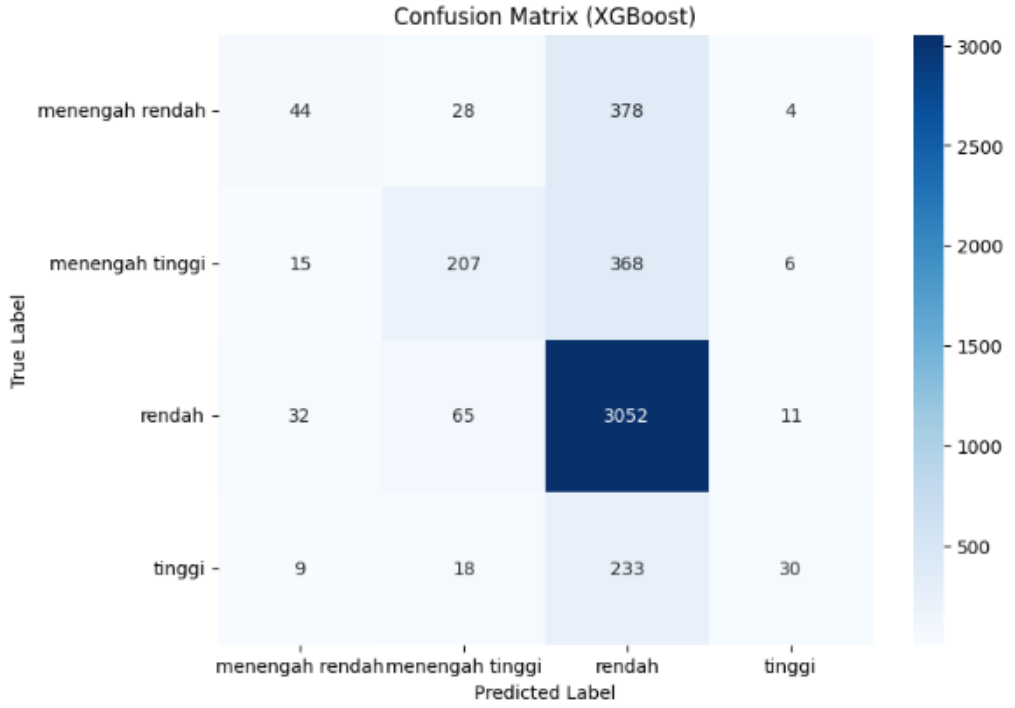
```
Classification Report (Random Forest):
                precision    recall  f1-score   support

menengah rendah       0.32      0.17      0.22       454
menengah tinggi       0.60      0.53      0.56       596
        rendah        0.80      0.90      0.85      3160
        tinggi        0.37      0.21      0.27       290

      accuracy                           0.73      4500
     macro avg        0.52      0.45      0.47      4500
  weighted avg        0.70      0.73      0.71      4500


Overall Accuracy (Random Forest): 0.73
```

**Gambar 6.** *Classification Report Random Forest*

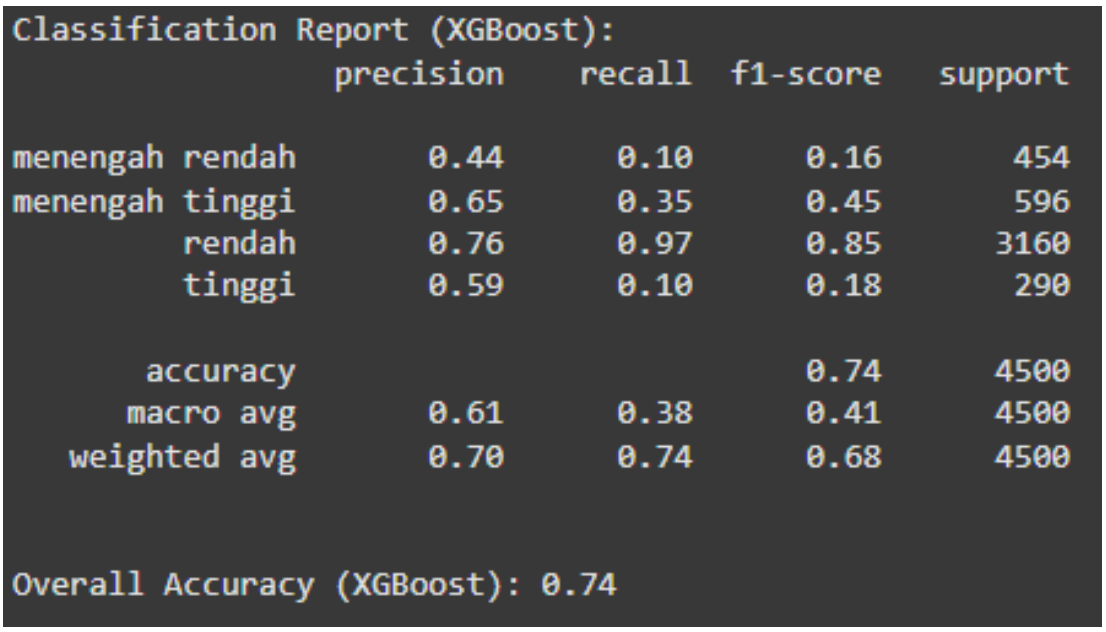
Keterangan Gambar:

1. Kinerja terbaik model Random Forest terdapat pada kelas "Rendah", dengan precision 0.80 dan recall 0.90, serta F1-score tinggi (0.85). Hal ini menunjukkan model sangat akurat dalam mengenali data dengan risiko rendah.
2. Kinerja paling lemah terjadi pada kelas "Menengah Rendah" dan "Tinggi", di mana F1-score sangat rendah (0.22 dan 0.27). Ini berarti banyak data dari dua kelas tersebut salah klasifikasi.
3. F1-score menengah tercapai di kelas "Menengah Tinggi", menunjukkan performa moderat.
4. Akurasi keseluruhan sebesar 73% menandakan model cukup baik, tetapi masih memiliki kelemahan dalam membedakan kelas-kelas risiko yang jumlah datanya lebih sedikit.

b. XGBoost menujukkan akurasi prediksi sebesar 0.74




**Gambar 7.** *Confusion Matrix XGBoost*



```
Classification Report (XGBoost):
                precision    recall  f1-score   support

menengah rendah       0.44      0.10      0.16       454
menengah tinggi       0.65      0.35      0.45       596
        rendah        0.76      0.97      0.85      3160
        tinggi        0.59      0.10      0.18       290

      accuracy                            0.74      4500
     macro avg        0.61      0.38      0.41      4500
  weighted avg        0.70      0.74      0.68      4500


Overall Accuracy (XGBoost): 0.74
```

**Gambar 8.** *Classification Report XGBoost*

Keterangan Gambar:

1. Kinerja terbaik model XGBoost terjadi pada kelas "Rendah", dengan nilai:
   a) Precision: 0.76
   b) Recall: 0.97
   c) F1-score: 0.85
   Ini menunjukkan bahwa model sangat baik dalam mengenali data risiko rendah, hampir tidak pernah melewatkan kasus (recall tinggi).
2. Kinerja buruk ditemukan pada kelas "Menengah Rendah" dan "Tinggi", yang hanya memiliki recall 0.10. Ini berarti model gagal mengenali sebagian besar data di kelas ini.
3. Kelas "Menengah Tinggi" memiliki performa menengah, lebih baik dari dua kelas sebelumnya namun masih jauh dari optimal.
4. Meskipun akurasi keseluruhan naik menjadi 74% (lebih tinggi dari Random Forest), ini sebagian besar disebabkan oleh dominasi kelas "Rendah" (3160 dari 4500 data), sehingga model lebih "berfokus" pada kelas ini.

**Tabel 1.** Perbandingan Random Forest dan XGBoost

| ***Aspek*** | ***Random Forest*** | ***XGBoost*** |
|---|---|---|
| *Akurasi* | 0.73 | 0.74 |
| *F1-score Kelas "Rendah"* | 0.85 | 0.85 |
| *F1-score Kelas Risiko Lain* | Lebih merata | Cenderung timpang |
| *Macro F1-score* | 0.47 | 0.41 |
| *Kesimpulan Umum* | Lebih seimbang antar kelas | Lebih kuat di kelas dominan |

Model *XGBoost* sedikit lebih unggul dibandingkan *Random Forest* dalam hal akurasi dan kemampuan membedakan antara kategori risiko. Precision yang tinggi pada kelas "Rendah" dalam XGBoost menunjukkan model ini cenderung membuat prediksi yang benar saat memprediksi risiko rendah.

### 3.2 Pemilihan Model

*F1-Score* dalam klasifikasi merupakan metrik yang sangat penting karena mengevaluasi keseimbangan antara *precision* (ketepatan prediksi positif) dan *recall* (kemampuan model dalam mendeteksi semua kasus positif yang relevan). Metrik ini memberikan indikator tunggal yang menggambarkan akurasi prediksi model secara lebih menyeluruh, terutama dalam situasi data yang tidak seimbang antar kelas. Dalam penelitian ini, meskipun perbedaan nilai akurasi antara model *Random Forest* dan *XGBoost* relatif kecil, keputusan pemilihan model tidak hanya didasarkan pada akurasi keseluruhan semata, melainkan mempertimbangkan performa model pada masing-masing kelas risiko.

Hal ini sangat penting, mengingat konsekuensi kesalahan klasifikasi pada kategori risiko tinggi dapat berimplikasi serius, terutama jika sebuah usaha berisiko tinggi diklasifikasikan secara keliru sebagai usaha dengan risiko rendah. Kesalahan seperti ini dapat menyebabkan UMKM yang seharusnya menerima pelatihan keselamatan kerja justru terabaikan, sehingga membahayakan keselamatan pekerja dan kelangsungan usaha itu sendiri. Oleh karena itu, metrik evaluasi seperti precision dan recall menjadi faktor pertimbangan utama dalam menentukan model terbaik, karena memberikan gambaran yang lebih rinci terhadap keandalan model dalam mengidentifikasi kelas risiko yang berdampak tinggi. *XGBoost*, meskipun tidak unggul pada semua kelas, dipilih karena menunjukkan performa lebih stabil dalam mengenali kategori risiko rendah secara konsisten dan akurat, serta memiliki keunggulan dalam menghindari underfitting pada data skala besar.

### 3.3 Implementasi dan Aplikasi

Hasil klasifikasi yang diperoleh dari model pembelajaran mesin ini memiliki nilai strategis yang signifikan dalam mendukung proses pengambilan keputusan, khususnya bagi pemangku kebijakan di lingkungan pemerintah daerah atau instansi terkait. Dengan adanya informasi yang lebih terstruktur mengenai tingkat risiko dari masing-masing UMKM, pihak berwenang dapat menentukan skala prioritas dalam penyusunan dan pelaksanaan program pelatihan keselamatan dan kesehatan kerja (K3). UMKM yang tergolong berisiko tinggi atau menengah tinggi dapat lebih dahulu mendapatkan intervensi dalam bentuk pelatihan, sosialisasi, atau audit keselamatan.

Selain itu, pendekatan klasifikasi ini juga memungkinkan penerapan kebijakan berbasis data (data-driven policy) yang lebih objektif dan efisien, di mana alokasi sumber daya seperti tenaga pelatih, anggaran pelatihan, serta waktu pelaksanaan dapat diarahkan secara tepat sasaran. Model ini dapat pula diintegrasikan ke dalam dashboard monitoring dan sistem perizinan OSS, sehingga proses deteksi risiko dan rekomendasi intervensi dapat dilakukan secara otomatis dan real time. Dengan demikian, implementasi model tidak hanya berdampak pada efisiensi administratif, tetapi juga berkontribusi langsung pada peningkatan kualitas keselamatan kerja, pengurangan kecelakaan, dan peningkatan keberlanjutan usaha, terutama di sektor UMKM yang sering kali memiliki keterbatasan sumber daya.

## 4. KESIMPULAN

Penelitian ini merupakan implementasi model klasifikasi risiko yang cocok untuk memberikan rekomendasi kepada pelaku usaha Mikro, Kecil, dan Menengah (UMKM) dalam pemberian program pelatihan keselamatan kerja dengan menggunakan pendekatan pembelajaran mesin (machine learning), khususnya melalui algoritma Random Forest dan XGBoost dengan metodologi CRISP-DM. Latar belakang dari penelitian ini adalah rendahnya kesadaran dan penerapan keselamatan kerja di sektor UMKM, terutama pada sektor-sektor dengan potensi risiko tinggi. Dengan memanfaatkan data dari Dinas XYZ yang diperoleh melalui sistem OSS, kami membangun model prediktif berbasis data untuk mengidentifikasi UMKM yang berpotensi mengalami risiko tinggi dan oleh karena itu harus menjadi prioritas dalam program pelatihan keselamatan kerja.

Melalui proses pelatihan dan pengujian model terhadap dataset yang terdiri dari lebih dari 150.000 entri UMKM dengan pembagian 15.000 data training dan 137.646 data testing, ditemukan bahwa kedua algoritma memiliki kinerja klasifikasi yang cukup baik. Random Forest menghasilkan akurasi sebesar 0.73, sedangkan XGBoost menghasilkan akurasi yang sedikit lebih tinggi yaitu 0.74. Selain dari akurasi, evaluasi juga dilakukan terhadap metrik precision, recall, dan F1-score. Hasil evaluasi ini menunjukkan bahwa XGBoost tidak hanya unggul dalam akurasi, tetapi juga lebih konsisten dalam memprediksi kategori risiko, khususnya pada kelas 'Rendah' yang memiliki precision tinggi.

Dengan demikian, dapat disimpulkan bahwa XGBoost merupakan algoritma yang lebih tepat digunakan dalam klasifikasi risiko UMKM berdasarkan data yang digunakan dalam penelitian ini. Implementasi metodologi CRISP-DM memastikan bahwa seluruh proses penelitian dilakukan secara terstruktur, dari pemahaman bisnis hingga deployment, dengan pendekatan iteratif yang memungkinkan perbaikan berkelanjutan. Ke depan, model ini dapat terus dikembangkan dengan menambahkan fitur-fitur tambahan yang lebih representatif, seperti sejarah kecelakaan kerja, jenis bahan baku yang digunakan, atau riwayat inspeksi keselamatan, serta dapat diintegrasikan

ke dalam sistem informasi pemerintah untuk memperkuat kebijakan berbasis data dalam meningkatkan keselamatan kerja di sektor UMKM.

## UCAPAN TERIMAKASIH

Kami menyampaikan rasa terima kasih yang sebesar-besarnya kepada Dinas XYZ atas bantuan dan kerja samanya dalam menyediakan data Usaha Mikro, Kecil, dan Menengah (UMKM) yang diperoleh melalui sistem Online Single Submission (OSS). Ketersediaan data yang lengkap dan relevan dari instansi tersebut merupakan fondasi penting yang memungkinkan penelitian ini dapat terlaksana dengan baik serta menghasilkan analisis yang mendalam dan bermakna. Kami juga mengucapkan terima kasih kepada dosen pembimbing yang telah memberikan bimbingan, arahan, serta kritik konstruktif sepanjang proses penyusunan penelitian ini. Dukungan akademik yang diberikan sangat membantu dalam memperkuat landasan metodologis serta memperjelas fokus penelitian.

## REFERENCES

[1] 'Undang-undang (UU) No. 1 Tahun 1970 Tentang Keselamatan Kerja'.

[2] R. Rst, R. Yulistria, E. P. Handayani, and S. Nursanty, 'PENGARUH KESELAMATAN DAN KESEHATAN KERJA (K3) TERHADAP PRODUKTIVITAS KERJA KARYAWAN', JURNAL SWABUMI, vol. 9, no. 2, 2021.

[3] 'UNDANG-UNDANG REPUBLIK INDONESIA NOMOR 20 TAHUN 2008 TENTANG USAHA MIKRO, KECIL, DAN MENENGAH'.

[4] P. N. Pemerinta et al., 'PERATURAN PEMERINTAH REPUBLIK INDONESIA NOMOR 7 TAHUN 2O2I TENTANG KEMUDAHAN, PELINDUNGAN, DAN PEMBERDAYAAN KOPERASI DAN USAHA MIKRO, KECIL, DAN MENENGAH'.

[5] Leo Breiman, 'Random Forests', Jan. 2021.

[6] Pete Chapman et al., CRISP-DM 1.0 Step-by-step data mining guide. Daimler Chrysler, 1999.

[7] R. Wirth and J. Hipp, 'CRISP-DM: Towards a Standard Process Model for Data Mining', 2000.

[8] A. Ananda Surya, D. Rizki Darmawan, and A. Solichin, 'Prediksi Kapabilitas Calon Debitur Menggunakan Analisis Data Machine Learning Dengan Metode Random Forest', doi: 10.33364/algoritma/v.22-1.1929.

[9] M. R. Muttaqin and M. Defriani, 'Algoritma K-Means untuk Pengelompokan Topik Skripsi Mahasiswa', ILKOM Jurnal Ilmiah, vol. 12, no. 2, pp. 121–129, Aug. 2020, doi: 10.33096/ilkom.v12i2.542.121-129.

[10] P. R. Sihombing and I. F. Yuliati, 'Penerapan Metode Machine Learning dalam Klasifikasi Risiko Kejadian Berat Badan Lahir Rendah di Indonesia', MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, vol. 20, no. 2, pp. 417–426, May 2021, doi: 10.30812/matrik.v20i2.1174.

[11] Adhelia Nurfira Rachmi, 'IMPLEMENTASI METODE RANDOM FOREST DAN XGBOOST PADA KLASIFIKASI CUSTOMER CHURN', 2020.

[12] M. Dava Maulana et al., 'ALGORITMA XGBOOST UNTUK KLASIFIKASI KUALITAS AIR MINUM', 2023. [Online]. Available: https://www.kaggle.com/datasets/adityak

[13] Intan Permata and Esther Sorta Mauli Nababan, 'Application Of Game Theory In Determining Optimum Marketing Strategy In Marketplace', JURNAL RISET RUMPUN MATEMATIKA

DAN ILMU PENGETAHUAN ALAM, vol. 2, no. 2, pp. 65–71, Jul. 2023, doi: 10.55606/jurrimipa.v2i2.1336.