

Penerapan Pembelajaran Mesin untuk Prediksi Keputusan Investasi pada UMKM

Arvyno Pranata Limahardja

Sistem dan Teknologi Informasi

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

Bandung, Indonesia

18222007@std.stei.itb.ac.id

David Dewanto

Sistem dan Teknologi Informasi

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

Bandung, Indonesia

18222027@std.stei.itb.ac.id

Ricky Wijaya

Sistem dan Teknologi Informasi

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

Bandung, Indonesia

18222043@std.stei.itb.ac.id

Nicolaas Heru Dreandachrista

Sistem dan Teknologi Informasi

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

Bandung, Indonesia

18222050@std.stei.itb.ac.id

Mohamad Maulana Firdaus Ramadhan

Sistem dan Teknologi Informasi

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

Bandung, Indonesia

18222140@std.stei.itb.ac.id

Ringkasan—Penelitian ini bertujuan untuk mengidentifikasi faktor-faktor penentu keberhasilan proposal investasi Usaha Mikro, Kecil, dan Menengah (UMKM) dalam konteks Tujuan Pembangunan Berkelanjutan (SDG) 9. Dengan menggunakan studi kasus dari dataset "Shark Tank US", penelitian ini menerapkan pendekatan pemodelan prediktif yang menggabungkan data terstruktur (finansial, demografi) dengan data tekstual dari deskripsi bisnis yang diproses menggunakan embedding Sentence-BERT. Berbagai algoritma klasifikasi dievaluasi secara komparatif, di mana model LightGBM (LGBM) yang telah melalui optimasi hiperparameter ekstensif menggunakan Optuna dipilih sebagai model final. Model ini berhasil mencapai akurasi sekitar 74.4% pada data uji. Hasil penelitian ini memberikan implikasi praktis bagi UMKM dalam menyusun strategi penawaran dan bagi investor sebagai alat penyaringan awal yang efisien.

Index Terms—Pemodelan Prediktif, Pendanaan UMKM, Machine Learning, Shark Tank, SDG 9

I. PENDAHULUAN

A. Latar Belakang

Tujuan Pembangunan Berkelanjutan (SDG) nomor 9 menekankan pentingnya membangun industri, inovasi, dan infrastruktur yang berkelanjutan. Akses terhadap pendanaan adalah salah satu fondasi utama untuk mencapai tujuan ini, terutama bagi Usaha Mikro, Kecil, dan Menengah (UMKM) yang merupakan pendorong vital inovasi dan pertumbuhan ekonomi. Namun, UMKM seringkali menghadapi tantangan dalam mendapatkan modal untuk berkembang.

Untuk mengatasi masalah ini, penelitian ini menggunakan studi kasus dari acara "Shark Tank US". Dataset ini sangat relevan karena sebagian besar pesertanya adalah pelaku UMKM dan menyajikan proses pengambilan keputusan investasi oleh investor berpengalaman. Dataset ini memuat informasi yang kaya, mencakup identitas perusahaan, latar belakang pemilik, dan detail penawaran investasi, yang memungkinkan analisis

mendalam mengenai faktor-faktor penentu keberhasilan sebuah proposal investasi.

B. Rumusan Masalah

Keputusan investasi dalam konteks *venture capital* seperti Shark Tank merupakan proses yang kompleks, dipengaruhi oleh dua dimensi utama: kualitas proposal bisnis itu sendiri dan faktor subjektif dari para investor yang membuat keputusan. Setiap investor diketahui memiliki preferensi atau "bias" terhadap sektor industri, model bisnis, atau bahkan tipe pendiri tertentu. Oleh karena itu, penelitian ini bertujuan untuk mengurai kedua dimensi tersebut.

Berdasarkan latar belakang tersebut, penelitian ini difokuskan untuk menjawab dua pertanyaan utama:

- 1) **Profil Perusahaan Ideal:** Perusahaan seperti apa yang memiliki probabilitas tertinggi untuk mendapatkan investasi? Ini mencakup analisis terhadap faktor-faktor seperti sektor industri, struktur penawaran finansial (valuasi dan ekuitas), serta karakteristik pendiri.
- 2) **Bias dan Preferensi Investor:** Bagaimana bias dan preferensi unik dari masing-masing investor dapat diidentifikasi?

C. Tujuan Penelitian

Secara spesifik, tujuan dari penelitian ini adalah:

- Menganalisis karakteristik UMKM yang menarik bagi investor, sekaligus memodelkan preferensi dan bias investasi dari setiap "shark" berdasarkan riwayat keputusan mereka.
- Membangun sebuah model prediktif yang mampu mengintegrasikan kedua sisi informasi tersebut—profil perusahaan dan preferensi investor—untuk menghasilkan pre-

diksi kelayakan investasi yang lebih akurat dan kontekstual.

- Memberikan jawaban yang mendalam pada pertanyaan "perusahaan seperti apa yang akan diinvestasikan" dengan menyediakan wawasan yang bergantung pada *siapa* investor yang ada di dalam panel.

D. Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

- **Bagi UMKM:** Memberikan panduan strategis mengenai karakteristik bisnis yang menarik bagi investor sekaligus membantu mereka menargetkan tipe investor yang paling sesuai dengan profil bisnisnya, sehingga meningkatkan peluang pendanaan.
- **Bagi Investor:** Menyediakan kerangka kerja berbasis data untuk memahami pola keputusan di pasar dan membantu menyaring peluang investasi yang paling sesuai dengan tesis dan preferensi investasi mereka.
- **Bagi Ekosistem Ekonomi:** Menciptakan pemahaman yang lebih transparan mengenai dinamika "kecocokan" antara UMKM dan investor, mendorong terciptanya *win-win solution* yang pada akhirnya mendukung industrialisasi berkelanjutan sejalan dengan SDG 9.

E. Batasan Masalah

Penelitian ini memiliki beberapa batasan sebagai berikut:

- Sumber data terbatas pada acara Shark Tank di Amerika Serikat, dari musim 1 hingga 16.
- Fokus analisis adalah pada momen pengambilan keputusan investasi dan tidak mencakup analisis kinerja atau keberhasilan perusahaan setelah mendapatkan pendanaan.
- Terdapat potensi bias seleksi yang melekat pada dataset, karena hanya *pitch* bisnis yang dinilai layak oleh tim produksi yang ditayangkan dan masuk ke dalam dataset.

II. METODE PENELITIAN

Metodologi penelitian ini dirancang sebagai serangkaian eksperimen yang terstruktur untuk membangun dan mengevaluasi model prediksi. Pendekatan ini mencakup tiga tahap eksperimen utama untuk mengidentifikasi arsitektur pemodelan yang paling efektif, diikuti oleh tahap optimasi dan evaluasi final.

A. Akuisisi dan Pra-pemrosesan Data

Penelitian ini dimulai dengan menggabungkan dua dataset dari Kaggle, yaitu "**Shark Tank US dataset**" [1] dan "**Shark Tank Companies**" [2]. Proses penggabungan dilakukan untuk memperkaya data dengan deskripsi naratif perusahaan, menggunakan nama perusahaan dan pengusaha sebagai kunci gabungan yang telah melalui proses pembersihan dan pemetaan manual. Dari data gabungan, sebanyak 450 baris data yang paling lengkap dipilih untuk pemodelan. Dimana sebanyak 232 data merupakan *deal*.

1) **Seleksi Fitur (Feature Selection):** Dari puluhan fitur yang tersedia setelah penggabungan, dilakukan proses seleksi fitur yang cermat. Tujuannya adalah untuk menggunakan hanya informasi yang relevan dan tersedia *sebelum* keputusan investasi dibuat, serta untuk mencegah terjadinya kebocoran data (*data leakage*). Beberapa kategori fitur tidak diikuti sertakan dalam pemodelan:

- **Fitur Hasil Kesepakatan:** Fitur seperti Total Deal Amount, Investment Amount Per Shark, dan semua kolom yang terkait dengan detail investasi spesifik dari para investor (Barbara Corcoran Investment Amount, dll.) dihilangkan. Fitur-fitur ini baru ada setelah kesepakatan terjadi dan penggunaannya akan menyebabkan kebocoran data.
- **Fitur Target:** Kolom *deal* dan *Got Deal* adalah variabel target yang akan diprediksi, sehingga secara alami tidak termasuk dalam set fitur input.
- **Fitur Identitas dan Indeks:** Fitur unik seperti *title*, *Startup Name*, *entrepreneurs*, *website*, dan nomor episode/*pitch* dihilangkan karena tidak memberikan pola yang dapat digeneralisasi dan berisiko menyebabkan *overfitting*.

Berdasarkan seleksi tersebut, set fitur final dipilih untuk merepresentasikan tiga aspek inti dari sebuah proposal investasi: (1) **Data Finansial Penawaran** (*askedfor*, *valuation*, dll.), (2) **Karakteristik Perusahaan dan Pendiri** (*category*, *Industry*, *Pitchers Gender*, dll.), (3) **Konteks Panel Investor** (misalnya *Barbara Corcoran Present*), dan (4) **Deskripsi Naratif Bisnis** (*description*).

2) **Pemrosesan Fitur Terpilih:** Data yang telah diseleksi kemudian melalui tahap pra-pemrosesan akhir, di mana fitur dibagi menjadi dua jenis:

- **Data Terstruktur:** Fitur-fitur numerik dinormalisasi menggunakan *StandardScaler*, sementara fitur-fitur kategorikal (seperti *Industry* dan *Pitchers Gender*) dikonversi menjadi representasi numerik menggunakan *OneHotEncoder*.
- **Data Teks:** Kolom *description* dibersihkan dari tag HTML, URL, dan karakter non-standar untuk persiapan proses ekstraksi fitur menggunakan teknik NLP.

B. Desain Eksperimen Pemodelan

Untuk menemukan arsitektur model terbaik, tiga eksperimen utama dirancang sebagai berikut.

1) **Eksperimen 1: Analisis Kontribusi Data Teks:** Eksperimen ini bertujuan untuk mengukur dampak penambahan informasi tekstual terhadap performa model.

- **Prosedur:** Kinerja sebuah model klasifikasi akan dievaluasi dalam dua skenario. Pertama, model akan dilatih hanya menggunakan data terstruktur. Kedua, model akan dilatih menggunakan gabungan data terstruktur dan data teks. Perbandingan hasil dari kedua skenario ini akan menentukan apakah fitur teks memberikan nilai tambah prediktif.

2) *Eksperimen 2: Perbandingan Teknik Representasi Teks (Embedding)*: Jika data teks terbukti bermanfaat, eksperimen ini akan menentukan metode terbaik untuk merepresentasikan-nya secara numerik. Tiga teknik akan dievaluasi:

- **TF-IDF**: Sebagai metode dasar berbasis frekuensi kata.
- **Word2Vec**: Untuk menangkap hubungan semantik antar kata.
- **Sentence-BERT**: Model transformer untuk menghasilkan embedding yang memahami konteks kalimat secara utuh.

Setiap teknik embedding akan diuji dengan arsitektur model yang sama untuk membandingkan mana yang menghasilkan performa terbaik.

3) *Eksperimen 3: Perbandingan Algoritma Klasifikasi*:

Pada tahap ini, beberapa keluarga algoritma klasifikasi akan diuji untuk mengidentifikasi mana yang paling cocok untuk dataset ini. Set fitur terbaik dari eksperimen sebelumnya akan digunakan untuk melatih dan membandingkan model-model berikut:

- **RandomForestClassifier**
- **XGBClassifier**
- **LGBMClassifier**
- **CatBoostClassifier**

Perbandingan ini akan didasarkan pada metrik performa awal untuk memilih algoritma yang paling menjanjikan.

C. Optimasi dan Evaluasi Model

Setelah serangkaian eksperimen di atas, metodologi akan ditutup dengan tahap optimasi dan evaluasi.

- **Optimasi Hiperparameter**: Algoritma yang menunjukkan kinerja terbaik dari Eksperimen 3 akan dioptimalkan lebih lanjut. Proses ini menggunakan pustaka Optuna untuk secara otomatis mencari kombinasi hiperparameter yang dapat memaksimalkan performa model.
- **Evaluasi Final**: Model akhir yang telah dioptimasi kemudian akan dievaluasi secara menyeluruh pada data uji. Kinerjanya akan dianalisis menggunakan metrik standar seperti Classification Report (yang mencakup *precision*, *recall*, *f1-score*) dan Confusion Matrix untuk mengukur kemampuan prediktifnya secara komprehensif.

III. HASIL DAN VISUALISASI

Bagian ini menyajikan hasil dari serangkaian eksperimen pemodelan yang telah dilakukan. Visualisasi data difokuskan pada perbandingan kinerja antar model dan evaluasi mendalam terhadap model final yang terpilih.

A. Hasil Perbandingan Model Eksperimental

Tiga eksperimen utama telah dilakukan untuk menemukan kombinasi arsitektur model dan representasi fitur yang paling efektif. Tabel I merangkum hasil akurasi dari setiap kombinasi yang diuji pada data validasi.

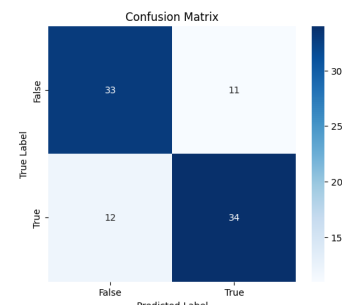
Dari hasil pada Tabel I, dapat ditarik beberapa kesimpulan awal. Pertama, penambahan fitur teks secara konsisten memberikan peningkatan performa dibandingkan dengan model yang hanya menggunakan data terstruktur. Kedua, kombinasi fitur **Terstruktur + BERT** menunjukkan hasil yang paling

unggul di antara semua set fitur, menandakan bahwa embedding berbasis transformer mampu menangkap informasi paling relevan dari deskripsi perusahaan. Di antara para algoritma, **LightGBM** menunjukkan potensi tertinggi.

B. Evaluasi Model Final (LGBM)

Berdasarkan hasil eksperimen awal dan proses optimasi hiperparameter, model **LightGBM (LGBM)** dipilih sebagai model final. Alasan pemilihan LGBM adalah karena keseimbangan antara performa yang kuat dan efisiensi waktu komputasi selama proses tuning. Kinerja model final ini kemudian dievaluasi secara mendalam pada data uji.

1) *Confusion Matrix*: Gambar 1 menyajikan *confusion matrix* dari hasil prediksi model final pada data uji. Matrix ini memvisualisasikan performa model dalam mengklasifikasikan setiap kelas, yaitu *True* (mendapat kesepakatan) dan *False* (tidak mendapat kesepakatan).



Gambar 1. Confusion Matrix Model Final (LGBM)

Dari matriks tersebut, terlihat bahwa model berhasil memprediksi dengan benar **33** kasus *False* (True Negative) dan **34** kasus *True* (True Positive).

2) *Laporan Klasifikasi*: Tabel II merinci metrik performa model final. Laporan ini mencakup *precision*, *recall*, dan *f1-score* untuk setiap kelas, serta akurasi keseluruhan.

Model final mencapai akurasi sebesar 74%. Nilai F1-score yang seimbang antara kedua kelas (0.73 untuk *False* dan 0.76 untuk *True*) menunjukkan bahwa model memiliki kemampuan yang baik dalam memprediksi kedua kemungkinan hasil tanpa bias yang signifikan terhadap salah satu kelas.

C. Analisis Kepentingan Fitur (Feature Importance)

Untuk memahami faktor-faktor yang memengaruhi prediksi model, dilakukan analisis menggunakan SHAP (*SHapley Additive exPlanations*). Agar interpretasi tetap fokus pada variabel yang mudah dipahami secara bisnis, analisis SHAP ini secara spesifik dijalankan dengan **mengecualikan fitur-fitur hasil rekayasa teknik** seperti vektor dari embedding teks (BERT) dan kolom-kolom hasil *one-hot encoding*.

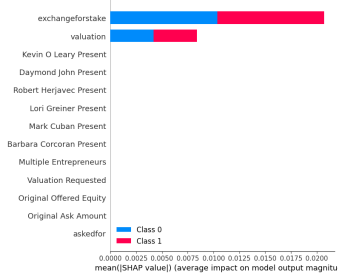
Dengan demikian, analisis ini bertujuan untuk melihat bagaimana model memanfaatkan fitur-fitur terstruktur asli dalam membuat keputusan. Hasilnya, yang ditunjukkan pada Gambar 2, mengilustrasikan kontribusi dari setiap fitur terstruktur tersebut.

Tabel I
HASIL METRIK PERFORMA DARI EKSPERIMEN

Set Fitur	Algoritma	Akurasi	F1-Score (Macro)	Precision (Macro)	Recall (Macro)
Hanya Terstruktur	RandomForest	0.4889	0.4866	0.4875	0.4876
	XGBoost	0.5222	0.5150	0.5208	0.5198
	LightGBM	0.4222	0.4150	0.4167	0.4200
	CatBoost	0.4778	0.4772	0.4772	0.4773
Terstruktur + TF-IDF	RandomForest	0.5111	0.5109	0.5119	0.5119
	XGBoost	0.4222	0.4211	0.4225	0.4234
	LightGBM	0.4444	0.4442	0.4449	0.4452
	CatBoost	0.5222	0.5222	0.5228	0.5227
Terstruktur + Word2Vec	RandomForest	0.5667	0.5662	0.5663	0.5662
	XGBoost	0.5444	0.5430	0.5438	0.5435
	LightGBM	0.5111	0.5101	0.5104	0.5104
	CatBoost	0.4778	0.4762	0.4766	0.4768
Terstruktur + BERT	RandomForest	0.6000	0.5968	0.6007	0.5983
	XGBoost	0.6000	0.5968	0.6007	0.5983
	LightGBM	0.6111	0.6087	0.6116	0.6097
	CatBoost	0.5444	0.5375	0.5442	0.5420

Tabel II
LAPORAN KLASIFIKASI MODEL FINAL (LGBM)

Kelas	Precision	Recall	F1-Score	Support
False (Tidak Deal)	0.76	0.70	0.73	44
True (Deal)	0.73	0.78	0.76	46
Accuracy			0.74	90
Macro Avg	0.75	0.74	0.74	90
Weighted Avg	0.75	0.74	0.74	90



Gambar 2. SHAP Summary Plot yang difokuskan pada Fitur Terstruktur Asli

IV. INTERPRETASI HASIL

Bagian ini menguraikan makna di balik angka dan visualisasi yang disajikan pada bagian sebelumnya. Interpretasi ini bertujuan untuk menerjemahkan hasil kuantitatif menjadi wawasan kualitatif mengenai faktor-faktor penentu keberhasilan investasi UMKM.

A. Interpretasi Hasil Eksperimen

Hasil dari serangkaian eksperimen pemodelan memberikan beberapa wawasan kunci mengenai data dan arsitektur model yang paling efektif.

- **Pentingnya Konteks Naratif (Data Teks):** Hasil pada Tabel I menunjukkan lompatan akurasi yang signifikan ketika fitur teks ditambahkan ke dalam model. Model

yang hanya menggunakan data terstruktur memiliki performa yang rendah (akurasi sekitar 44%-48%), yang menyiratkan bahwa angka-angka finansial saja (seperti valuasi atau jumlah modal yang diminta) tidak cukup untuk membuat prediksi yang baik. Performa model meningkat drastis saat informasi dari kolom ‘description’ disertakan. Hal ini mengindikasikan bahwa narasi, masalah yang diselesaikan, dan proposisi nilai dari sebuah bisnis—yang terkandung dalam deskripsi teks—merupakan faktor krusial dalam pertimbangan investor.

- **Keunggulan Embedding Kontekstual (BERT):** Di antara tiga metode representasi teks yang diuji, **BERT** secara konsisten memberikan performa terbaik dengan selisih yang cukup jauh. Ini berarti keputusan investor tidak hanya dipengaruhi oleh kata kunci (seperti yang ditangkap TF-IDF) atau makna kata secara individual (Word2Vec), tetapi oleh pemahaman kontekstual dari keseluruhan deskripsi bisnis. Kemampuan BERT untuk memahami nuansa kalimat dan hubungan antar kata terbukti paling efektif dalam mengekstrak sinyal prediktif dari data naratif.

B. Analisis Faktor Prediktif Utama

Analisis SHAP yang terfokus (Gambar 2) memberikan pemahaman mendalam tentang logika bisnis yang telah dipelajari oleh model:

- **Faktor Penentu Utama: Struktur Kesepakatan Finansial:** Analisis menegaskan bahwa fitur yang paling berpengaruh adalah **exchangeforstake** (persentase ekuitas yang ditawarkan) dan **valuation** (valuasi perusahaan). Ini adalah temuan yang sangat intuitif, yang menandakan bahwa model telah berhasil belajar seperti seorang investor rasional, di mana fundamental penawaran finansial menjadi pertimbangan utama.

- **Hipotesis Preferensi Investor Tidak Terbukti:** Temuan paling signifikan dari analisis ini adalah **terbantahkannya hipotesis awal** bahwa preferensi subjektif investor (yang diwakili oleh fitur kehadiran mereka) akan mempengaruhi model. Data SHAP secara kuantitatif menunjukkan bahwa fitur seperti *Kevin O Leary Present* dan kehadiran investor lainnya memiliki dampak nol. Model tidak menemukan pola yang dapat digeneralisasi dari kehadiran investor dan memilih untuk mengabaikan informasi ini sepenuhnya.

C. Interpretasi Kinerja Model Final (LGBM)

Model final LGBM yang telah dioptimasi menunjukkan kinerja yang menjanjikan dan memberikan wawasan praktis bagi calon investor maupun UMKM.

- **Akurasi dan Keseimbangan Model:** Dengan akurasi keseluruhan sebesar **74%**, model ini mampu memprediksi hasil proposal investasi dengan benar untuk sekitar 3 dari 4 kasus pada data uji. Lebih penting lagi, nilai F1-score yang seimbang untuk kelas *True* (0.76) dan *False* (0.73) seperti yang terlihat pada Tabel II, menunjukkan bahwa model tidak bias terhadap salah satu kelas. Model ini sama baiknya dalam mengidentifikasi perusahaan yang berpotensi mendapatkan kesepakatan maupun yang tidak, menjadikannya alat bantu keputusan yang andal.
- **Analisis Kesalahan (Confusion Matrix):** Analisis pada Gambar 1 memberikan pemahaman tentang jenis kesalahan yang dibuat model:
 - **False Positive (13 kasus):** Model memprediksi sebuah perusahaan akan mendapatkan investasi, padahal kenyataannya tidak. Dari perspektif investor, ini adalah representasi dari **risiko investasi**, di mana model mungkin merekomendasikan sebuah proposal yang sebenarnya kurang layak.
 - **False Negative (10 kasus):** Model memprediksi sebuah perusahaan tidak akan mendapat investasi, padahal kenyataannya berhasil. Ini adalah representasi dari **peluang yang terlewatkan (*missed opportunity*)**, di mana investor bisa kehilangan kesempatan berinvestasi pada bisnis yang potensial.
- **Implikasi Praktis untuk Pengambilan Keputusan:**
 - **Recall (78% untuk kelas *True*):** Menunjukkan bahwa model ini adalah alat penyaringan (*screening tool*) yang baik. Ia mampu mengidentifikasi hampir 4 dari 5 perusahaan yang sesungguhnya berhasil mendapatkan kesepakatan.
 - **Precision (73% untuk kelas *True*):** Memberikan "pemeriksaan realitas". Ketika model memprediksi *Deal*, ada kemungkinan 73% prediksi tersebut benar. Ini menekankan bahwa meskipun model memberikan sinyal positif, proses *due diligence* lebih lanjut oleh investor tetap sangat diperlukan.

V. REKOMENDASI

Berdasarkan analisis dan interpretasi model yang telah divalidasi, bagian ini menguraikan beberapa implikasi praktis serta

saran perbaikan lanjutan yang berfokus pada pengayaan data dan pengembangan teknik pemodelan.

A. Implikasi Hasil dan Penerapan Praktis

Model yang dikembangkan berhasil mengidentifikasi faktor-faktor fundamental dalam keputusan investasi, sehingga memiliki implikasi yang kuat dan dapat ditindaklanjuti.

- **Bagi Pelaku UMKM:** Hasil penelitian ini memberikan panduan strategis yang jelas. Temuan bahwa faktor penentu utama adalah struktur penawaran finansial (*exchange for stake* dan *valuation*) menggarisbawahi pentingnya bagi UMKM untuk fokus pada valuasi yang realistis dan penawaran ekuitas yang menarik. Di samping itu, pentingnya data teks juga menunjukkan bahwa narasi bisnis yang kuat tetap krusial untuk menarik perhatian awal investor.
- **Bagi Investor:** Model ini dapat dikembangkan menjadi sistem pendukung keputusan (*Decision Support System*) yang objektif. Investor dapat menggunakannya sebagai alat penyaringan awal untuk mengidentifikasi proposal yang paling menjanjikan berdasarkan fundamental finansialnya, sehingga memungkinkan proses *due diligence* menjadi lebih efisien dan terfokus pada aspek kualitatif bisnis.
- **Bagi Ekosistem Pendanaan:** Dengan menyediakan pendekatan berbasis data yang transparan, penelitian ini berkontribusi pada tujuan SDG 9 dengan cara menjembatani kesenjangan antara UMKM inovatif yang membutuhkan modal dan investor yang mencari peluang investasi yang layak.

B. Saran untuk Perbaikan dan Pengembangan Lanjutan

Meskipun model yang dihasilkan menunjukkan performa yang baik, terdapat beberapa area pengembangan yang berfokus pada pemodelan untuk meningkatkan kapabilitas dan akurasinya di masa depan.

- **Penambahan Fitur Finansial dan Operasional:** Untuk membangun model yang lebih komprehensif, disarankan untuk memperkaya dataset dengan fitur-fitur kuantitatif yang lebih detail, seperti data historis pendapatan (*revenue*), margin keuntungan (*gross margin*), biaya akuisisi pelanggan (*customer acquisition cost*), atau jumlah pengguna aktif jika relevan. Penambahan fitur-fitur ini akan memungkinkan model untuk melakukan analisis fundamental yang lebih mendalam.
- **Validasi dan Kontekstualisasi Menggunakan Data Lokal:** Untuk meningkatkan relevansi praktis model di Indonesia, langkah krusial berikutnya adalah melakukan validasi dan pelatihan ulang menggunakan dataset yang berisi data UMKM dan keputusan investasi dari pasar lokal. Hal ini penting karena dinamika pasar, valuasi, dan perilaku investor di Indonesia mungkin memiliki karakteristik yang unik.

PUSTAKA

- [1] T. Thirumani, "Shark Tank US Dataset," Kaggle, 2025. [Online]. Tersedia: <https://www.kaggle.com/datasets/thirumani/shark-tank-us-dataset>.
- [2] U. T. Pedersen, "Shark Tank Companies," Kaggle, 2017. [Online]. Tersedia: <https://www.kaggle.com/datasets/ulrikthgepedersen/shark-tank-companies>.