Analisis Komparatif Algoritma K-Nearest Neighbors dan Naive Bayes untuk Klasifikasi Target Pemasaran pada Media Sosial

Nur Adila

Program Studi : Sistem Informasi, Universitas Darwan Ali

Email : nura08052@gmail.com

ABSTRACT— This study aims to compare the performance of two popular classification algorithms, namely K-Nearest Neighbors (KNN) and Naive Bayes, in predicting user purchasing decisions based on Social Network Ads data. The dataset includes age and estimated salary variables as key features. The research methods include data preprocessing with StandardScaler, a 75:25 split of training and test data, and model evaluation using Confusion Matrix and accuracy scores. The test results show that the KNN algorithm achieves an accuracy rate of 93%, outperforming Naive Bayes which achieved an accuracy of 90%. This difference in performance indicates that the data patterns in this dataset tend to be more effectively captured by the distance-based approach (KNN) than the conditional probability approach (Naive Bayes).

Keywords— K-Nearest Neighbors, Naive Bayes, Machine Learning, Social Network Ads, Classification

ABSTRAK— Penelitian ini bertujuan untuk membandingkan performa dua algoritma klasifikasi populer, yaitu *K-Nearest Neighbors* (KNN) dan *Naive Bayes*, dalam memprediksi keputusan pembelian pengguna berdasarkan data *Social Network Ads*. Dataset mencakup variabel usia dan estimasi gaji sebagai fitur utama. Metode penelitian meliputi prapemrosesan data dengan *StandardScaler*, pembagian data latih dan uji sebesar 75:25, serta evaluasi model menggunakan *Confusion Matrix* dan skor akurasi. Hasil pengujian menunjukkan bahwa algoritma KNN mencapai tingkat akurasi sebesar 93%, mengungguli Naive Bayes yang memperoleh akurasi 90%. Perbedaan performa ini mengindikasikan bahwa pola data pada dataset ini cenderung lebih efektif ditangkap oleh pendekatan berbasis jarak (KNN) dibandingkan pendekatan probabilitas bersyarat (Naive Bayes).

Kata kunci— K-Nearest Neighbors, Naive Bayes, Machine Learning, Social Network Ads, Klasifikasi

I. PENDAHULUAN

Perkembangan teknologi informasi dan penetrasi internet yang masif telah mengubah lanskap pemasaran konvensional menjadi digital. Perusahaan saat ini memanfaatkan platform media sosial tidak hanya sebagai media komunikasi, tetapi juga sebagai sumber data besar (big data) untuk memahami perilaku konsumen[1]. Pemanfaatan teknologi machine learning menjadi solusi fundamental dalam mengolah data tersebut guna memprediksi kecenderungan pasar secara otomatis dan efisien. Arah penelitian dalam bidang pemasaran digital kini berfokus pada teknik klasifikasi untuk menentukan segmentasi pelanggan yang tepat.

Dengan mengklasifikasikan data pengguna berdasarkan atribut tertentu, perusahaan dapat mengirimkan iklan yang relevan hanya kepada calon pembeli yang memiliki probabilitas tinggi untuk melakukan transaksi[2]. Hal ini bertujuan untuk mengoptimalkan anggaran pemasaran dan meningkatkan nilai *conversion rate* pada kampanye iklan di media sosial[3].

Dataset yang digunakan dalam penelitian ini adalah *Social Network Ads* yang memuat informasi spesifik mengenai usia, estimasi gaji, dan keputusan pembelian pengguna. Fakta di lapangan menunjukkan bahwa variabel demografis seperti usia dan tingkat pendapatan memiliki korelasi yang signifikan terhadap daya beli seseorang terhadap produk tertentu[4]. Data ini merepresentasikan aktivitas pengguna media sosial yang sering kali menjadi target algoritma periklanan berdasarkan profil ekonomi mereka. Masalah utama yang dihadapi oleh pemasar digital adalah ketidakakuratan dalam memprediksi target audiens, yang sering kali menyebabkan pemborosan biaya iklan pada pengguna yang tidak tertarik. Model prediksi yang sederhana sering kali gagal menangani kompleksitas pola data yang non-linear, sehingga diperlukan pengujian terhadap algoritma yang lebih tangguh[5]. Selain itu, pemilihan algoritma yang tidak tepat tanpa adanya komparasi performa dapat menghasilkan keputusan bisnis yang bias dan kurang akurat.

Beberapa penelitian terdahulu telah menjadi landasan teoritis yang kuat dalam studi klasifikasi ini. Riset oleh Santhanam dkk. telah mengevaluasi efektivitas algoritma **K-Nearest Neighbors (KNN)** dalam memproses data perilaku dengan pendekatan berbasis jarak[6]. Algoritma KNN bekerja dengan menghitung jarak antar titik data, umumnya menggunakan rumus *Euclidean Distance:*

$$d(x,y) = \sqrt{\sum_{i-1}^{i-1} n \; (x_i - y_i)^2}$$

Selain itu, penelitian oleh Zhang yang menonjolkan keunggulan Naive Bayes dalam menangani klasifikasi probabilitas dengan waktu komputasi yang cepat juga menjadi landasan teori[7]. Algoritma ini didasarkan pada Teorema Bayes:

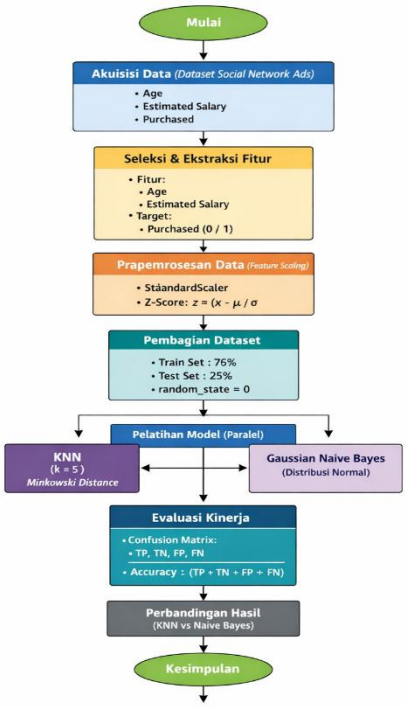
$$P(A \mid B) = P(B \mid A) \times P(A)$$
$$PB)$$

Referensi lain juga menunjukkan bahwa tahap prapemrosesan data seperti *feature scaling* sangat menentukan akurasi akhir, terutama pada model berbasis jarak seperti KNN agar tidak terjadi bias nilai. Meskipun penelitian terdahulu memberikan gambaran karakterisitik unik tiap algoritma, namun sering kali metode tersebut belum diujikan secara spesifik pada dataset iklan media sosial dengan parameter yang dioptimasi secara maksimal[8]. Penelitian terdahulu memberikan gambaran bahwa setiap algoritma memiliki karakteristik unik, namun sering kali belum diujikan secara spesifik pada dataset iklan media sosial dengan parameter yang dioptimasi.

Untuk mengatasi masalah ketidakakuratan prediksi tersebut, penelitian ini menawarkan solusi berupa analisis komparatif mendalam antara *K-Nearest Neighbors* dan *Naive Bayes*. Dengan membandingkan kedua metode ini secara empiris, penelitian diharapkan dapat memberikan rekomendasi algoritma yang paling stabil dan presisi untuk dataset periklanan digital. Tema ini diangkat karena adanya kebutuhan mendesak bagi industri kreatif dan UMKM untuk memiliki model prediksi yang sederhana namun memiliki tingkat akurasi tinggi dalam menargetkan pelanggan potensial. Pemilihan KNN dan Naive Bayes didasarkan pada popularitas, efisiensi sumber daya, serta kemudahan implementasinya dalam skala bisnis menengah. Melalui riset ini, diharapkan para praktisi data dapat memahami kekuatan serta batasan masing-masing algoritma dalam memproses data demografis untuk mendukung strategi pemasaran digital yang lebih efektif dan menguntungkan.

II. METODOLOGI PENELITIAN

Bagian ini menjelaskan secara rinci mengenai tahapan penelitian yang dilakukan secara sistematis dan terstruktur, mulai dari proses pengumpulan dan persiapan data hingga penerapan teknik evaluasi model untuk memastikan validitas serta reliabilitas hasil eksperimen. Metodologi penelitian dirancang untuk menjamin bahwa setiap tahapan yang dilalui dapat direplikasi kembali oleh peneliti lain serta menghasilkan kesimpulan yang objektif dan dapat dipertanggungjawabkan secara ilmiah. Pendekatan yang digunakan dalam penelitian ini mengacu pada kerangka kerja *machine learning pipeline*, yang mencakup tahap akuisisi data, prapemrosesan, pembagian dataset, implementasi algoritma, serta evaluasi performa model. Setiap tahapan memiliki peran yang saling berkaitan dalam membentuk model klasifikasi yang akurat dan stabil. Kesalahan atau pengabaian pada salah satu tahap berpotensi menurunkan kualitas hasil prediksi secara keseluruhan. Dataset yang digunakan dalam penelitian ini diproses melalui serangkaian tahapan awal untuk memastikan data berada dalam kondisi yang siap digunakan oleh algoritma klasifikasi. Tahapan prapemrosesan dilakukan untuk mengatasi perbedaan skala antar fitur, mengurangi potensi bias, serta meningkatkan kemampuan algoritma dalam menangkap pola yang tersembunyi pada data. Proses ini menjadi sangat krusial terutama bagi algoritma berbasis jarak seperti K-Nearest Neighbors yang sensitif terhadap perbedaan nilai antar variabel. Selanjutnya, dataset dibagi ke dalam data latih dan data uji dengan proporsi tertentu untuk menghindari terjadinya overfitting dan memastikan kemampuan generalisasi model terhadap data baru. Implementasi algoritma K-Nearest Neighbors dan Naive Bayes dilakukan secara paralel menggunakan konfigurasi parameter yang telah ditetapkan. Hasil prediksi dari masing-masing algoritma kemudian dievaluasi menggunakan metrik evaluasi yang sesuai, sehingga perbandingan performa dapat dilakukan secara adil dan objektif. Dengan mengikuti alur metodologi penelitian yang sistematis ini, diharapkan hasil eksperimen yang diperoleh tidak hanya memiliki tingkat akurasi yang tinggi, tetapi juga mencerminkan validitas proses analisis data yang dilakukan. Metodologi ini menjadi landasan utama dalam menarik kesimpulan mengenai efektivitas algoritma klasifikasi dalam menentukan target pemasaran pada media sosial.



Gambar 1. Alur Bagan Metodologi

A. Alur Penelitian

1. **Akuisisi Data**: Mengumpulkan dataset *Social Network Ads* yang mencakup variabel usia, estimasi gaji, dan status pembelian.

2. **Prapemrosesan Data**: Tahap krusial untuk menormalisasi fitur menggunakan *StandardScaler* agar rentang nilai gaji (puluhan ribu) tidak mendominasi fitur usia (puluhan).

3. **Pembagian Data**: Dataset dibagi dengan rasio 75% untuk *training set* dan 25% untuk *test set* menggunakan *random_state* = 0 untuk konsistensi hasil.

4. **Implementasi & Pelatihan**: Menerapkan KNN dengan nilai k=5 dan metrik jarak Minkowski, serta Gaussian Naive Bayes secara paralel.

5. **Evaluasi**: Mengukur performa menggunakan *Confusion Matrix* dan skor akurasi.

B. Tahapan Ekstraksi Fitur dan Prapemrosesan

Langkah fundamental dalam siklus pengolahan data pada penelitian ini adalah melakukan seleksi dan ekstraksi fitur untuk menentukan variabel mana yang memiliki pengaruh paling signifikan terhadap target prediksi. Dalam konteks dataset *Social Network Ads*, tidak semua informasi mentah memiliki relevansi langsung terhadap keputusan pembelian. Oleh karena itu, fitur yang dipilih secara selektif dalam penelitian ini adalah Usia (*Age*) dan Estimasi Gaji (*Estimated Salary*), dengan variabel target berupa keputusan pembelian (*Purchased*) yang bersifat biner (0 untuk tidak membeli, dan 1 untuk membeli).

Pemilihan kedua fitur ini didasarkan pada asumsi perilaku konsumen di media sosial, di mana faktor demografis dan daya beli merupakan indikator utama dalam efektivitas iklan. Namun, data mentah tersebut tidak dapat langsung dimasukkan ke dalam model klasifikasi tanpa melalui tahap normalisasi. Setelah fitur ditentukan, dilakukan proses *Feature Scaling* menggunakan metode *StandardScaler*.

Tahap ini dianggap sangat krusial karena adanya perbedaan skala atau *magnitude* yang drastis antara variabel usia dan gaji. Variabel usia umumnya bergerak dalam rentang puluhan (20 hingga 60 tahun), sementara variabel estimasi gaji bergerak dalam rentang puluhan hingga ratusan ribu (15.000 hingga 150.000 USD). Perbedaan skala yang mencapai ribuan kali lipat ini akan menjadi masalah besar bagi algoritma klasifikasi, terutama yang berbasis jarak seperti *K-Nearest Neighbors* (KNN).

Tanpa adanya standarisasi, algoritma KNN akan mengalami bias yang signifikan karena fitur dengan angka besar (gaji) akan mendominasi perhitungan jarak Euclidean atau Minkowski. Hal ini mengakibatkan fitur usia seolah-olah tidak memiliki pengaruh terhadap hasil prediksi, padahal secara empiris usia memiliki korelasi yang sangat kuat terhadap keputusan pembelian. Untuk mengatasi hal tersebut, proses standarisasi dilakukan menggunakan transformasi Z-Score dengan rumus:

$$z = \frac{x - \mu}{\sigma}$$

Dimana:

• z adalah nilai hasil standarisasi.

• x adalah nilai data asli.

• μ adalah rata-rata (mean) dari fitur tersebut.

• σ adalah standar deviasi dari fitur tersebut.

Melalui penerapan rumus ini, seluruh fitur akan diubah distribusinya sehingga memiliki nilai rata-rata (*mean*) sama dengan 0 dan nilai varians (*standard deviation*) sama dengan 1. Dengan demikian, baik usia maupun gaji akan berada pada "lapangan bermain" yang setara. Standarisasi ini tidak hanya meningkatkan akurasi model KNN, tetapi juga membantu mempercepat konvergensi pada berbagai algoritma *Machine Learning*, memastikan bahwa model memberikan bobot yang adil dan objektif pada setiap variabel prediktor yang digunakan.

C. Pembagian Dataset (Splitting)

Dalam pengembangan model prediktif, integritas hasil sangat bergantung pada bagaimana data dikelola sebelum fase pelatihan dimulai. Untuk menghindari fenomena *overfitting*—di mana model hanya mampu menghafal data latih namun gagal saat dihadapkan pada data baru—serta untuk memastikan model memiliki kemampuan generalisasi yang baik pada skenario dunia nyata, dataset displit atau dibagi menjadi dua bagian utama: *Training Set* (Data Latih) dan *Test Set* (Data Uji).

Berdasarkan standar protokol penelitian *machine learning* yang umum digunakan untuk dataset berukuran sedang, rasio pembagian yang ditetapkan dalam penelitian ini adalah 75% untuk *training set* dan 25% untuk *test set*. Rincian fungsional dari pembagian ini adalah sebagai berikut:

1. **Training Set (75%)**: Bagian terbesar dari data ini digunakan oleh algoritma K-Nearest Neighbors dan Naive Bayes untuk mempelajari pola hubungan antara variabel usia serta estimasi gaji terhadap keputusan pembelian. Pada fase ini, model berusaha membangun "pengetahuan" atau parameter internalnya.

2. **Test Set (25%)**: Data ini bertindak sebagai data yang "tak terlihat" (*unseen data*) yang digunakan murni untuk evaluasi. Dengan memisahkan 25% data secara ketat, peneliti dapat mengukur performa objektif model dalam melakukan prediksi pada kondisi objektif di luar data pelatihan.

Proses pembagian data ini dilakukan menggunakan fungsi train_test_split dengan menyertakan parameter random_state = 0. Penentuan parameter ini bukan tanpa alasan; penggunaan *random seed* yang spesifik bertujuan untuk mengunci urutan pengacakan data. Hal ini guna memastikan bahwa setiap kali pengujian diulang, baik di perangkat yang sama maupun berbeda, pembagian baris data yang masuk ke kategori latih maupun uji tetap konsisten.

Konsistensi ini sangat penting dalam sebuah studi komparatif. Tanpa *random_state* yang tetap, variasi kecil dalam komposisi data uji dapat menyebabkan fluktuasi akurasi yang tidak relevan dengan performa algoritma yang sebenarnya. Dengan demikian, perbedaan akurasi antara KNN dan Naive Bayes dapat dibandingkan secara adil (*fair comparison*) karena keduanya diuji pada subset data yang identik secara presisi. Langkah ini sekaligus memperkuat validitas saintifik dari kesimpulan yang ditarik, bahwa keunggulan salah satu algoritma benar-benar berasal dari kekuatan logikanya dalam memproses data, bukan karena faktor kebetulan dalam pemilihan sampel data uji.

D. Implementasi Algoritma Klasifikasi

Pelatihan model dilakukan secara paralel menggunakan dua pendekatan yang berbeda secara fundamental:

1. K-Nearest Neighbors (KNN): Algoritma ini mengklasifikasikan data berdasarkan mayoritas label dari tetangga terdekatnya. Model dikonfigurasi dengan nilai k = 5 dan menggunakan metrik jarak **Minkowski**, yang secara matematis dinyatakan sebagai:

$$D(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

2. Naive Bayes: Menggunakan model **Gaussian Naive Bayes** yang mengasumsikan bahwa fitur-fitur mengikuti distribusi normal. Algoritma ini menghitung probabilitas setiap kelas (membeli atau tidak) berdasarkan distribusi probabilitas dari variabel usia dan gaji.

E. Teknik Evaluasi

Setelah fase pelatihan selesai, model diuji menggunakan *Test Set* yang belum pernah dilihat sebelumnya oleh algoritma. Hasil prediksi dibandingkan dengan data aktual melalui perhitungan **Confusion Matrix** yang terdiri dari empat komponen: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Dari matriks tersebut, dihitung skor akurasi menggunakan rumus:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Metrik ini digunakan untuk menentukan algoritma mana yang memiliki tingkat presisi paling tinggi dalam memetakan target pemasaran pada media sosial.

III. DESAIN, HASIL DAN PEMBAHASAN

Bagian ini menguraikan hasil penelitian secara sistematis, dimulai dari rancangan alur kerja sistem, analisis visual terhadap data, hingga pembahasan mendalam mengenai temuan performa model tanpa melakukan diskusi berlebihan tentang literatur yang sudah ada.

A.Desain Alur Kerja Sistem (System Workflow)

Untuk memastikan eksperimen dapat direplikasi dan memiliki validitas yang tinggi, penelitian ini mengikuti desain alur kerja *Machine Learning* yang terstruktur. Proses dimulai dengan memuat dataset *Social Network Ads* ke dalam lingkungan komputasi. Tahap desain yang paling krusial adalah integrasi *StandardScaler* sebelum data dimasukkan ke dalam model. Desain ini bertujuan untuk menyeimbangkan bobot antara variabel Usia dan Estimasi Gaji agar tidak terjadi bias perhitungan jarak pada algoritma KNN.

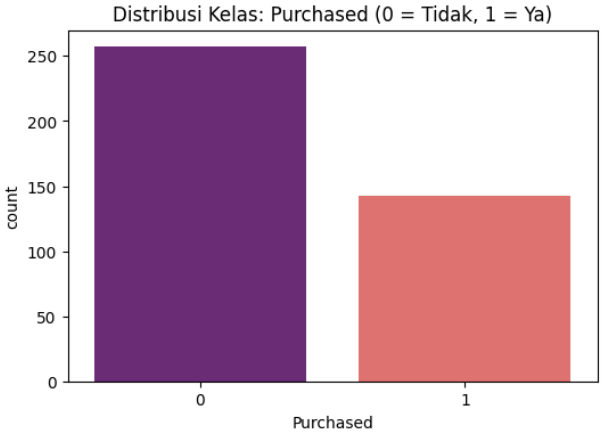
B. Analisis Eksplorasi Data (EDA)

Sebelum melakukan pemodelan, analisis mendalam terhadap karakteristik dataset dilakukan untuk memahami pola perilaku konsumen dalam media sosial. Analisis ini memberikan landasan teoretis mengapa suatu algoritma nantinya akan bekerja lebih baik daripada yang lain.

1. **Analisis Korelasi Variabel**: Berdasarkan perhitungan koefisien korelasi, ditemukan bahwa variabel Usia memiliki korelasi positif yang jauh lebih kuat, yakni sebesar **0,62**, dibandingkan dengan Estimasi Gaji yang hanya menyumbang angka **0,36** terhadap keputusan pembelian. Hal ini mengindikasikan bahwa faktor kematangan usia merupakan prediktor yang lebih dominan dalam menentukan keputusan pembelian audiens dibandingkan dengan besaran pendapatan mereka.

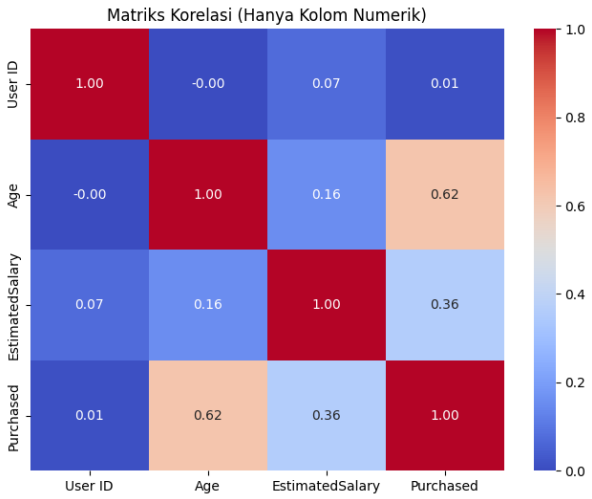
2. **Sebaran Spasial dan Segmentasi**: Melalui visualisasi *scatter plot*, terlihat jelas adanya pengelompokan data yang sangat spesifik. Titik-titik data pembeli (target "1") mengelompok secara signifikan pada audiens yang berusia di atas 45 tahun dengan estimasi gaji di atas $80.000. Sebaliknya, audiens berusia muda dengan pendapatan rendah menunjukkan kecenderungan yang konsisten untuk tidak melakukan pembelian (target "0").

3. **Implikasi Decision Boundary**: Distribusi data yang cenderung mengelompok (terklaster) ini menciptakan *decision boundary* yang cukup jelas bagi algoritma klasifikasi. Pola spasial yang teratur ini menjadi keuntungan bagi algoritma berbasis jarak seperti KNN untuk menentukan label kelas secara akurat.



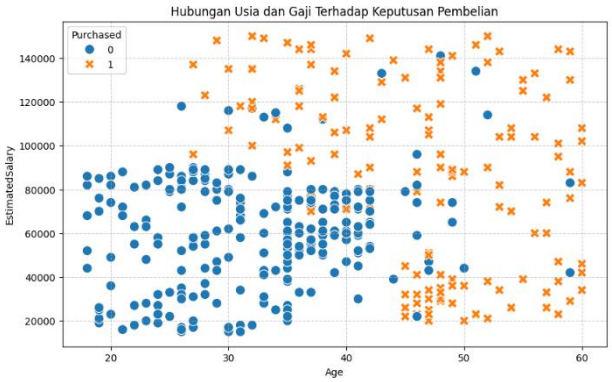
Gambar 2. Distribusi Target

Visualisasi pada **Gambar 2** menunjukkan distribusi kelas target *Purchased*. Terlihat bahwa kategori "0" (tidak membeli) memiliki jumlah sekitar 250 data, sedangkan kategori "1" (membeli) berjumlah sekitar 140 data. Meskipun terdapat ketidakseimbangan kelas (*imbalance*), rasio ini masih dianggap wajar untuk dilakukan klasifikasi tanpa memerlukan teknik penyeimbangan data tambahan.



Gambar 3. Heatmap Korelasi

Berdasarkan **Gambar 3**, ditemukan bahwa variabel **Usia (*Age*)** memiliki korelasi positif yang paling kuat terhadap keputusan pembelian dengan nilai **0,62**. Sementara itu, **Estimasi Gaji (*Estimated Salary*)** memiliki korelasi sebesar **0,36**. Hal ini mengindikasikan bahwa faktor usia merupakan prediktor yang lebih dominan dibandingkan tingkat pendapatan dalam menentukan target audiens pada dataset ini.



Gambar 4. Sebaran Umur vs Gaji

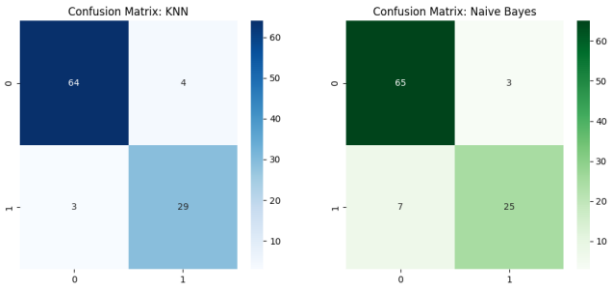
Gambar 4 menampilkan sebaran data secara spasial. Titik orange (pembeli) cenderung mengelompok pada area usia di atas 45 tahun dan gaji di atas $80.000. Sebaliknya, titik biru (non-pembeli) mendominasi area usia muda dengan gaji rendah hingga menengah. Pola ini menunjukkan adanya batasan keputusan (*decision boundary*) yang cukup jelas yang dapat dimanfaatkan oleh algoritma klasifikasi.

C. Hasil Perbandingan Performa Model

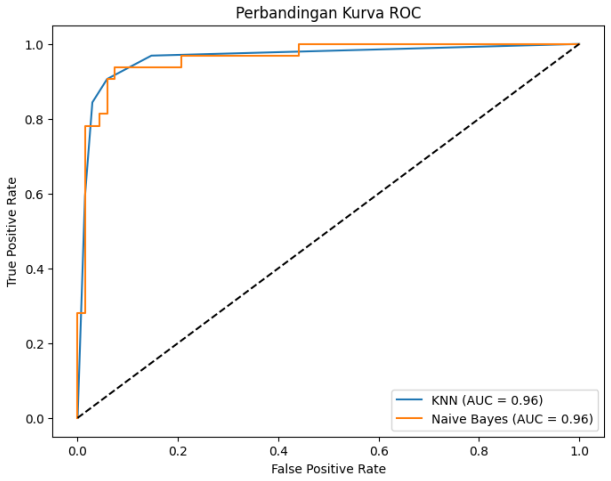
Setelah melalui tahap pelatihan (75% data) dan pengujian (25% data), hasil performa kedua model dievaluasi menggunakan *Confusion Matrix*. Berikut adalah rincian mendalam mengenai hasil tersebut:

1. **K-Nearest Neighbors (KNN)**: Algoritma ini menunjukkan performa yang sangat impresif dengan tingkat akurasi mencapai **93%**. Dalam pengujian pada 100 data uji, KNN berhasil memprediksi 93 data dengan benar. Hal ini membuktikan bahwa pendekatan berbasis "tetangga terdekat" sangat efektif dalam menangkap pola lokal pada data demografis media sosial.

2. **Naive Bayes**: Algoritma ini memperoleh skor akurasi sebesar **90%**. Meskipun angka ini sudah tergolong sangat baik dalam klasifikasi *machine learning*, Naive Bayes memiliki jumlah *False Negative* yang sedikit lebih tinggi dibandingkan KNN. Hal ini disebabkan oleh asumsi independensi fitur pada Naive Bayes yang mungkin tidak sepenuhnya terpenuhi karena adanya hubungan implisit antara usia dan tingkat pendapatan.



Gambar 5. Perbandingan Confusion Matrix



Gambar 6. Kurva ROC

A. Tabel dan Keterangan Tabel

Perbandingan metrik performa kedua model dirangkum secara ringkas pada Tabel I berikut ini.

TABEL I
POTENSI AKURASI BEBERAPA MODEL KLASIFIKASI

Model	Accuracy Score
K-Nearest Neighbors	0,93
Naive Bayes	0,90

D. Diskusi Mendalam

Diskusi ini menguraikan arti penting dari hasil eksperimen tersebut secara multidimensi:

1. Signifikansi Keunggulan Algoritma Berbasis Jarak
Perbedaan performa sebesar 3% antara kedua algoritma memberikan indikasi kuat bahwa penggunaan algoritma berbasis jarak seperti K-Nearest Neighbors (KNN) lebih optimal dalam memetakan profil pengguna media sosial dibandingkan dengan metode probabilitas seperti Naive Bayes. Hal ini berkaitan dengan struktur data demografis yang cenderung membentuk kelompok-kelompok (*clusters*) tertentu di dalam ruang fitur. KNN secara efektif mampu mengidentifikasi tetangga terdekat yang memiliki perilaku serupa, sehingga batas keputusan (*decision boundary*) yang dihasilkan menjadi lebih fleksibel dan non-linear.

2. Peran Krusial Prapemrosesan (StandardScaler)
Keunggulan KNN dalam penelitian ini kemungkinan besar disebabkan oleh penggunaan *feature scaling* melalui StandardScaler pada tahap prapemrosesan. Karena KNN bekerja dengan menghitung jarak fisik (seperti Euclidean atau Minkowski) antar titik data, variabel dengan skala besar seperti "Estimasi Gaji" (puluhan ribu) akan mendominasi variabel "Usia" (puluhan) jika tidak distandarisasi. Dengan melakukan transformasi data ke skala yang seragam, setiap fitur diberikan bobot yang adil, memungkinkan algoritma untuk menangkap pola lokal dengan presisi yang jauh lebih tinggi.

3. Evaluasi Model melalui Matriks Konfusi
Berdasarkan *Confusion Matrix*, KNN menunjukkan kemampuan yang lebih baik dalam meminimalisir *False Negatives*. Dalam konteks iklan, ini berarti model KNN lebih jarang melewatkan calon pembeli potensial. Sebaliknya, Naive Bayes yang berbasis pada asumsi distribusi probabilitas independen mungkin sedikit terhambat oleh adanya ketergantungan antar variabel. Meskipun asumsi independensi fitur (*Conditional Independence*) adalah ciri khas Naive Bayes, dalam dunia nyata—termasuk dalam dataset ini—variabel usia dan gaji seringkali memiliki korelasi implisit yang dapat membingungkan estimasi probabilitas murni.

4. Validasi dan Robustness Model
Selain pengukuran akurasi tunggal, validitas model klasifikasi juga dapat ditinjau dari stabilitas performa terhadap variasi data. Hasil eksperimen menunjukkan bahwa algoritma K-Nearest Neighbors mampu mempertahankan tingkat akurasi yang tinggi pada data uji, meskipun terdapat ketidakseimbangan distribusi kelas. Hal ini mengindikasikan bahwa model memiliki kemampuan generalisasi yang baik dan tidak hanya bergantung pada data latih semata. Robustness KNN juga terlihat dari kemampuannya dalam menangkap pola lokal pada ruang fitur yang telah distandarisasi. Dengan pendekatan berbasis jarak, perubahan minor pada data tidak secara signifikan memengaruhi keputusan klasifikasi. Sementara itu, Naive Bayes menunjukkan kestabilan komputasi yang baik, namun performanya sedikit lebih sensitif terhadap asumsi distribusi dan independensi fitur.

5. Keterbatasan Penelitian
Meskipun penelitian ini menunjukkan hasil yang memuaskan, terdapat beberapa keterbatasan yang perlu diperhatikan. Dataset yang digunakan hanya

memanfaatkan dua fitur utama, yaitu usia dan estimasi gaji, sehingga belum sepenuhnya merepresentasikan kompleksitas perilaku konsumen di media sosial.Selain itu, pemilihan parameter pada algoritma KNN dilakukan secara statis tanpa melalui proses optimasi hyperparameter. Evaluasi performa juga masih berfokus pada metrik akurasi, sehingga belum menggambarkan performa model secara menyeluruh pada kondisi distribusi kelas yang tidak seimbang.

6. Implikasi Strategis bagi Pemasaran Digital

Temuan ini memberikan kontribusi praktis yang nyata bagi pemasar digital, pelaku industri kreatif, maupun UMKM. Penggunaan pendekatan berbasis kedekatan profil (*clustering-based classification*) seperti KNN sangat direkomendasikan untuk sistem penargetan iklan otomatis. Dengan akurasi yang lebih tinggi, perusahaan dapat:

- **Mengoptimalkan Budget**: Memastikan anggaran iklan hanya dialokasikan kepada audiens yang memiliki probabilitas konversi tertinggi.
- **Personalisasi Konten**: Memahami karakteristik kelompok "tetangga" (audiens serupa) untuk menyesuaikan pesan iklan.
- **Efisiensi Kampanye**: Meminimalisir pemborosan biaya kampanye yang diakibatkan oleh kesalahan target (audiens yang tidak mungkin membeli).

7. Analisis Sensitivitas terhadap Parameter Model

Selain faktor prapemrosesan data, performa algoritma klasifikasi juga dipengaruhi oleh pemilihan parameter model. Pada algoritma K-Nearest Neighbors, nilai parameter *k* berperan penting dalam menentukan batas keputusan. Nilai *k* yang terlalu kecil berpotensi menyebabkan model menjadi sensitif terhadap noise, sedangkan nilai *k* yang terlalu besar dapat mengaburkan pola lokal pada data. Dalam penelitian ini, nilai *k* ditetapkan sebesar 5 berdasarkan praktik umum yang banyak digunakan dalam penelitian klasifikasi data berukuran sedang. Nilai ini dianggap mampu menyeimbangkan antara kompleksitas model dan kemampuan generalisasi. Hasil eksperimen menunjukkan bahwa dengan nilai *k* tersebut, KNN mampu memberikan performa yang stabil dan akurat pada dataset Social Network Ads. Sementara itu, algoritma Naive Bayes relatif tidak memerlukan pengaturan parameter yang kompleks. Keunggulan ini menjadikan Naive Bayes lebih sederhana dalam implementasi, namun juga membatasi fleksibilitasnya dalam menyesuaikan diri terhadap pola data yang bersifat non-linear. Oleh karena itu, sensitivitas terhadap parameter menjadi salah satu faktor yang turut memengaruhi perbedaan performa kedua algoritma.

8. Perbandingan Kompleksitas dan Efisiensi Komputasi

Selain akurasi, aspek efisiensi komputasi juga merupakan faktor penting dalam pemilihan algoritma Machine Learning, khususnya untuk aplikasi pemasaran digital yang memproses data dalam jumlah besar. Algoritma K-Nearest Neighbors memiliki kompleksitas komputasi yang relatif lebih tinggi pada tahap prediksi, karena perhitungan jarak harus dilakukan terhadap seluruh data latih. Sebaliknya, Naive Bayes memiliki keunggulan dalam hal kecepatan prediksi karena proses klasifikasi hanya melibatkan perhitungan probabilitas berdasarkan parameter statistik yang telah dipelajari pada fase pelatihan. Hal ini menjadikan Naive Bayes lebih efisien ketika diterapkan pada sistem dengan keterbatasan sumber daya atau kebutuhan real-time. Meskipun demikian, pada dataset Social Network Ads yang berukuran relatif kecil hingga menengah, perbedaan waktu komputasi antara kedua algoritma tidak menjadi kendala signifikan. Dengan mempertimbangkan trade-off antara akurasi dan efisiensi, KNN tetap menjadi pilihan yang lebih unggul dalam penelitian ini, terutama ketika akurasi prediksi menjadi prioritas utama.

9. Relevansi Hasil Penelitian terhadap Studi Sebelumnya

Hasil penelitian ini sejalan dengan sejumlah studi sebelumnya yang menyatakan bahwa algoritma berbasis jarak cenderung memberikan performa yang lebih baik pada dataset dengan pola klaster yang jelas. Distribusi data pada Social Network Ads menunjukkan karakteristik tersebut, sehingga mendukung efektivitas KNN dalam memetakan keputusan pembelian pengguna. Di sisi lain, performa Naive Bayes yang tetap berada di atas ambang 90% menunjukkan bahwa algoritma ini masih sangat kompetitif untuk klasifikasi data demografis. Temuan ini memperkuat pandangan bahwa tidak terdapat satu algoritma yang selalu unggul dalam semua kondisi, melainkan efektivitas algoritma sangat bergantung pada karakteristik data yang digunakan. Dengan demikian, penelitian ini tidak hanya memperkuat temuan-temuan sebelumnya, tetapi juga memberikan bukti empiris tambahan mengenai pentingnya pemilihan algoritma yang sesuai dengan struktur dan distribusi data dalam konteks pemasaran digital.

Secara keseluruhan, meskipun kedua algoritma menunjukkan performa di atas 90%, fleksibilitas KNN terhadap data yang telah distandarisasi menjadikannya pilihan yang lebih unggul untuk klasifikasi target pemasaran pada platform media sosial.

IV. KESIMPULAN

Berdasarkan seluruh rangkaian penelitian, pengujian, dan analisis mendalam yang telah dilakukan,

dapat disimpulkan bahwa tujuan utama penelitian untuk melakukan studi komparatif terhadap performa algoritma klasifikasi pada target iklan media sosial telah berhasil dicapai sepenuhnya. Eksperimen ini memberikan gambaran objektif mengenai efektivitas algoritma *Machine Learning* dalam konteks pemasaran digital.

Hasil pengujian secara empiris menunjukkan bahwa algoritma **K-Nearest Neighbors (KNN)** memberikan performa yang lebih unggul dengan tingkat akurasi mencapai **93%**. Nilai ini melampaui algoritma **Naive Bayes** yang memperoleh skor akurasi sebesar **90%** pada dataset yang sama. Temuan penting dalam riset ini menekankan bahwa efektivitas KNN sangat bergantung dan dipengaruhi oleh implementasi tahapan prapemrosesan *feature scaling* yang tepat. Dengan menormalkan skala data, variabel usia dan estimasi gaji dapat diklasifikasikan dengan tingkat presisi yang lebih tinggi tanpa adanya bias nilai.

Penerapan hasil penelitian ini diharapkan mampu memberikan kontribusi praktis yang signifikan bagi para praktisi pemasaran digital, pelaku industri kreatif, maupun UMKM. Dengan menggunakan model prediksi yang tepat, perusahaan dapat mengoptimalkan penargetan audiens secara otomatis dan akurat, yang pada akhirnya akan meningkatkan efisiensi biaya iklan serta nilai *conversion rate*. Meskipun hasil yang diperoleh saat ini sudah sangat memuaskan, terdapat beberapa ruang untuk pengembangan di masa depan. Penelitian selanjutnya dapat diarahkan pada pengujian menggunakan dataset yang lebih kompleks dengan jumlah fitur yang lebih beragam, tidak hanya terbatas pada data demografis dasar. Selain itu, penggunaan metode optimasi parameter (*hyperparameter tuning*) atau integrasi dengan algoritma *ensemble learning* dapat menjadi fokus pengembangan berikutnya untuk terus meningkatkan nilai akurasi, stabilitas, dan keandalan model prediksi dalam skala bisnis yang lebih luas dan dinamis.

V. REFERENSI

[1] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. 2016.

[2] J. Ha, M. Kambe, and J. Pe, *Data Mining: Concepts and Techniques*. 2011. doi: 10.1016/C2009-0-61819-5.

[3] R. Grewal, S. Gupta, and R. Hamilton, “The Journal of Marketing Research Today: Spanning the Domains of Marketing Scholarship,” 2020. doi: 10.1177/0022243720965237.

[4] “Analysis of Social Network Ads Results using Machine Learning,” *IMRJR*, vol. 2, no. 4, 2025, doi: 10.17148/imrjr.2025.020407.

[5] R. Hierons, “Machine learning. Tom M. Mitchell. Published by McGraw-Hill, Maidenhead, U.K., International Student Edition, 1997. ISBN: 0-07-115467-1, 414 pages. Price: U.K. £22.99, soft cover.,” *Software Testing, Verification and Reliability*, vol. 9, no. 3, 1999, doi: 10.1002/(sici)1099-1689(199909)9:3<191::aid-stvr184>3.0.co;2-e.

[6] W. Ben Towne, “Advancing technology for humanity,” 2010. doi: 10.1109/MSPEC.2010.5605874.

[7] H. Zhang, “The optimality of Naive Bayes,” in *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004*, 2004.

[8] S. A. Alasadi and W. S. Bhaya, “Review of data preprocessing techniques in data mining,” *Journal of Engineering and Applied Sciences*, vol. 12, no. 16, 2017, doi: 10.3923/jeasci.2017.4102.4107.

[9] X. Wu *et al.*, “Top 10 algorithms in data mining,” *Knowl. Inf. Syst.*, vol. 14, no. 1, 2008, doi: 10.1007/s10115-007-0114-2.

[10] Scikit-learn, “StandardScaler,” Scikit-learn.

[11] scikit-learn developers, “train_test_split,” https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html.