Inserch
Information System Research Journal

E-ISSN: 2775-4669
Volume 4, Nomor 2, Agustus 2024

Penerapan Algoritma Klasifikasi untuk Menentukan Segmentasi Pelanggan Berbasis Machine Learning

Novia Lestari[1], Julianti[2], M. Rafly Syahrul Fathan[3]
Prodi Sistem Informasi, Universitas Islam Negeri Imam Bonjol Padang[1,2,3]
Balai Gadang, Kecamatan Koto Tangah, Kota Padang, Sumatera Barat[1,2,3]
Noviia.lestari@uinib.ac.id[1], juli90433@gmail.com [2] , prapthorespect@gmail.com [3]

Received: 19 Aug 2024 | Revised: 04 Sep 2024
Accepted: 24 Sep 2024 | Published: 01 Oct 2024

Abstrak

Segmentasi pelanggan merupakan konsep pemasaran yang sangat penting dalam konteks relationship marketing yang dapat meningkatkan pemahaman tentang kebutuhan pelanggan yang lebih baik untuk menciptakan strategi pemasaran yang lebih efektif dan personal. Oleh karena itu dibutuhkan analisis yang tepat sesuai karakteristik yang ada untuk menentukan segmentasi pelanggan tersebut. Pada penelitian ini, untuk mengidentifikasi pola tersembunyi dan secara otomatis mengelompokkan pelanggan digunakan tiga algoritma klasifikasi yaitu Random Forest, Support Vector Machine (SVM), dan Gradient Boosting. Hasilnya menunjukkan bahwa segmentasi pelanggan berhasil membagi konsumen ke dalam empat kelompok yang berbeda: A, B, C, dan D, masing-masing dengan karakteristik yang unik. Model Gradient Boosting memiliki kinerja terbaik dalam validasi silang dengan akurasi 54%, meskipun semua model menunjukkan penurunan akurasi pada data uji. Hasil segmentasi pelanggan pada penelitian ini dapat memberikan wawasan yang berharga bagi perusahaan otomotif untuk merancang strategi pemasaran yang lebih tepat sasaran dan efektif untuk meningkatkan profitabilitas.

Kata kunci : Segmentasi, Support Vector Machine, Random Forest, Gradient Boosting

Abstract

Customer segmentation is a very important marketing concept in the context of relationship marketing that can improve understanding of customer needs to create a more effective and personalized marketing strategy. Therefore, an appropriate analysis is needed to determine customer segmentation according to the existing characteristics. In this research, three classification algorithms are used to identify hidden patterns and automatically segment customers: Random Forest, Support Vector Machine (SVM), and Gradient Boosting. The results show that customer segmentation successfully divides consumers into four different groups: A, B, C, and D, each with unique characteristics. The Gradient Boosting model has the best performance in cross-validation with 54% accuracy, although all models show a decrease in accuracy on test data. The customer segmentation results in this study can provide valuable insights for automotive companies to design more targeted and effective marketing strategies to increase profitability.

Keywords: Segmentation, Support Vector Machine, Random Forest, Gradient Boosting

I. PENDAHULUAN

Perkembangan industri otomotif semakin kompetitif berdampak pada kemampuan untuk memahami preferensi pelanggan sebagai strategi kunci untuk meningkatkan profit perusahaan [1]. Faktor pelanggan merupakan salah satu hal penting bagi perusahaan dalam bersaing dengan kompetitor lainnya, namun saat ini informasi mengenai karakteristik mereka belum cukup. Penelitian ini dilakukan untuk menganalisis karakteristik pelanggan sebagai dasar dalam penetapan segmentasi dan customer *profiling*. Segmentasi pelanggan memungkinkan perusahaan mengelompokkan konsumen ke dalam segmen-segmen yang lebih homogen berdasarkan karakteristik tertentu, sehingga strategi pemasaran dapat dirancang lebih tepat sasaran untuk meningkatkan profit perusahaan [2].

Pada pengelompokkan pelanggan sesuai kelas yang ada, teknik data mining seperti klasifikasi dapat digunakan [3]. Teknik klasifikasi dapat digunakan untuk mengelompokkan data ke dalam kategori atau kelas tertentu. Dalam klasifikasi,

model dilatih untuk mempelajari pola dalam data berdasarkan fitur-fitur yang ada, dengan tujuan untuk memprediksi label atau kelas dari data baru yang belum pernah dilihat [4]. Terdapat beberapa algoritma yang bisa digunakan dalam teknik klasifikasi seperti Random Forest, *Support Vector Machine* (SVM), dan Gradient Boosting.

Penelitian terdahulu membahas analisis dan penerapan Algoritma Support Vector Machine (SVM) dalam data mining untuk mendukung strategi promosi di Universitas PGRI Semarang. Dengan menggunakan metode SVM, dapat ditemukan pola-pola tersembunyi dalam data yang ada, sehingga strategi promosi yang lebih efisien dan efektif dapat diterapkan [5]. Penelitian lainnya membandingkan tiga algoritma untuk analisis sentiment terhadap stigma kesehatan mental dengan hasil pengujian model SVM performa terbaik dengan akurasi 86.11%. Kemudian model Random Forest dengan akurasi 82.55%, sedangkan model Naive bayes dengan akurasi sebesar 78,19% [6]. Sehingga dalam penelitian ini, ketiga algoritma ini dipilih karena kekuatan masing-masing dalam menghasilkan segmentasi yang akurat dan bermanfaat.

Penelitian ini terkait dengan sebuah perusahaan otomotif yang berencana memasuki pasar baru dengan produk yang sudah ada. Setelah melakukan riset pasar yang intensif, mereka menyimpulkan bahwa perilaku pasar baru tersebut mirip dengan pasar yang sudah ada. Di pasar yang sudah ada, tim penjualan telah mengklasifikasikan semua pelanggan ke dalam 4 segmen (A, B, C, D). Penelitian ini menggunakan pemodelan ketiga algoritma klasifikasi yaitu SVM, Gradient Boosting, dan Random Forest, kemudian dipilih satu model terbaik dengan akurasi tertinggi dari algoritma yang ada tersebut untuk memprediksi segmentasi yang tepat untuk pelanggan baru tersebut, sehingga dapat membatu perusahaan dalam menyusun strategi pemasaran yang lebih efektif.

II. TINJAUAN PUSTAKA

A. Data Mining

Data mining merupakan proses untuk menemukan pola dan informasi tersembunyi di dalam database, menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengidentifikasi pengetahuan dari basis data tersebut. Istilah lain yang sering digunakan secara sinonim dengan data mining adalah Knowledge Discovery in Database (KDD) [7].

Penelitian terdahulu membahas penerapan metode data mining berbasis clustering, khususnya algoritma k-means, untuk menilai status gizi balita di Puskesmas Kelapa Dua Tangerang. Hasil penelitian Data Mining Metode k-means clustering terbukti efektif dalam mengidentifikasi pola dan karakteristik status gizi balita, yang dapat digunakan untuk tindakan yang lebih tepat dalam menangani masalah gizi [8].

B. Random Forest

Random Forest adalah algoritma machine learning yang digunakan untuk pengklasifikasian data dalam jumlah besar. Algoritma ini merupakan gabungan dari beberapa pohon keputusan (decision tree) yang diperoleh melalui proses bagging atau bootstrap aggregating[9]. Untuk melakukan prediksi menggunakan metode Random Forest, hasil dari setiap pohon keputusan dikombinasikan dengan cara majority vote untuk klasifikasi dan rata-rata untuk regresi. Hasil akurasi yang dihasilkan Random forest terbukti bagus, kuat terhadap outliers dan noise, dan lebih cepat dibandingkan dengan bagging dan boosting [2].

Penelitian terdahulu membahas penerapan algoritma Random Forest untuk menghasilkan model prediksi hujan yang menghasilkan akurasi sebesar 71,09% dengan menggunakan teknik 10-fold cross validation dan akurasi sebesar 99,45% dengan teknik menggunakan seluruh data secara random [7].

C. Support Vector Machine

Support Vector Machine (SVM) merupakan salah satu metode klasifikasi yang sangat populer dan efisien dalam mengatasi berbagai permasalahan klasifikasi. SVM pertama kali diperkenalkan oleh Vladimir Vapnik pada tahun 1995, dan prinsip utamanya adalah mencari sebuah hyperplane atau batas pemisah terbaik yang memisahkan data ke dalam dua kelas dengan margin pemisah yang terbesar [6]. Margin ini adalah jarak terpendek antara titik data dari setiap kelas dengan hyperplane. Titik-titik data yang terletak di tepi margin tersebut disebut sebagai support vectors, dan merekalah yang menentukan posisi hyperplane yang optimal. Proses klasifikasi dalam SVM juga dapat diartikan sebagai usaha untuk menemukan hyperplane yang memisahkan dua kelompok data tersebut [10].

Metode SVM telah banyak digunakan pada penelitian salah satunya Perbandingan Hasil Klasifikasi Hujan Menggunakan Metode SVM dan Naïve Bayes Classification, dimana analisis SVM menggunakan model kernel Linier lebih baik

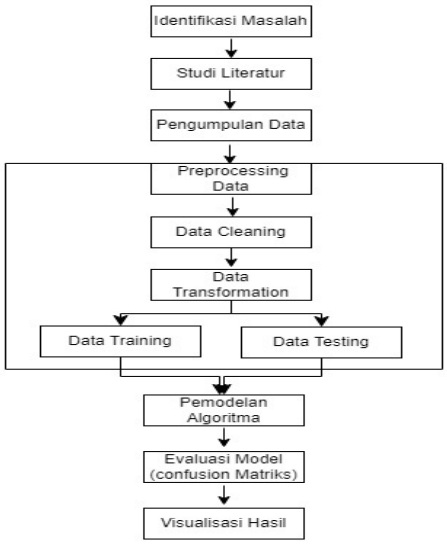
dalam memprediksi curah hujan dibandingkan metode NBC [11].

D. Gradient Boosting

Gradient Boosting adalah salah satu algoritma klasifikasi yang menggunakan pohon keputusan untuk membantu menyelesaikan permasalahan yang berkaitan dengan prediksi dan klasifikasi [12]. Dengan hanya menggunakan satu algoritma, gradient boosting dapat menyelesaikan kedua teknik tersebut sekaligus. Algoritma ini dikenal juga dengan algoritma pohon yang dapat menghindari overfitting dengan cara memperbaiki model secara iteratif dan menggabungkan prediksi dari beberapa model yang lebih sederhana untuk mencapai hasil yang lebih akurat [13].

III. METODE PENELITIAN

Berikut ini merupakan tahapan penelitian yang dilakukan dalam penelitian ini :



Gambar 1. Tahapan Penelitian

1. Identifikasi Masalah
   Masalah yang diteliti pada penelitian ini adalah bagaimana menentukan segmentasi pelanggan pada perusahaan otomotif.
2. Studi Literatur
   Tujuan dari studi literatur adalah untuk memahami perkembangan penelitian sebelumnya, mengidentifikasi gap atau celah dalam pengetahuan, serta membangun landasan teori dan konteks penelitian. Dalam proses ini, peneliti membaca secara kritis, mencatat poin-poin penting, dan mengorganisir informasi untuk mendukung argumen atau hipotesis penelitian [14].
3. Pengumpulan Data
   Data set yang digunakan dalam penelitian ini yaitu berisi informasi mengenai data pelanggan, seperti karakteristik demografis, preferensi, dan perilaku pembelian. Dataset terdiri dari 2.627 entri dengan 10 atribut dan 1 atribut target. Berikut ini adalah daftar dari dataset yang digunakan dalam penelitian ini.

Tabel 1. Daftar Atribut yang Digunakan

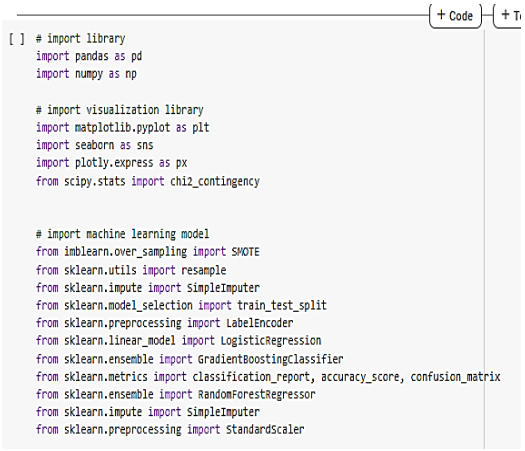
No	Attribut	Deskripsi
1	ID	Identifikasi unik yang dimiliki oleh setiap pelanggan
2	Gender	Menginformasikan jenis kelamin pelanggan
3	Ever_Married	Menginformasikan status pernikahan pelanggan
4	Age	Menginformasikan umur pelanggan
5	Graduated	Menginformasikan status pendidikan pelanggan
6	Profession	Menginformasikan jenis pekerjaan pelanggan
7	Work Experience	Menginformasikan pengalaman kerja pelanggan
8	Spending_Score	Menginformasikan pola pengeluaran pelanggan
9	Family_Size	Menginformasikan jumlah aggota keluarga
10	Var_1	Kategori anonym pelanggan
11	Segmentation	Segmen pelanggan yang menjadi target prediksi (A,B,C,D)

4. Preprocessing Data
   Setelah dataset diimpor, proses preprocessing data dilakukan untuk memahami struktur data, misalnya jumlah kolom, jenis data di setiap kolom (numerik, kategorikal), dan melihat apakah ada nilai yang hilang (missing values). Eksplorasi awal ini membantu mengidentifikasi potensi masalah atau pola dalam data sebelum melanjutkan ke tahap selanjutnya.
5. Cleaning Data
   Cleaning Data dilakukan untuk mempersiapkan data agar dapat diguakan dalam pemodelan algoritma. Berikut ini merupakan beberapa langkah yang dilakukan dalam tahapan cleaning data yaitu :
   a. Eksplorasi Struktur Data yaitu memahami bentuk dataset yang digunakan. Termasuk jumlah kolom, tipe data, serta melihat deskripsi statistik awal seperti distribusi data.
   b. Memeriksa nilai yang hilang (missing values) yaitu mengidentifikasi data yang hilang dan menentukan cara menanganinya, seperti menghapus baris atau kolom, atau mengisi kolom yang hilang dengan nilai dari rata-rata.

6. Transformasi Data
   Transformasi data melibatkan pengubahan format atau kondisi data dari satu bentuk ke bentuk lain, tanpa mengubah konten dataset. Bertujuan untuk memudahkan penggunaan data dalam analisis, integrasi data, dan pembuatan model [15]. Tranformasi data juga dapat diartikan sebagai pencarian fitur-fitur yang berguna untuk mempresentasikan data bergantung kepada goal yang ingin dicapai, seperti mengeksplorasi pola atau hubungan yang ada dalam dataset.
   Berikut ini merupakan langkah-langkah yang digunakan dalam tahapan ini:
   a. Visualisasi data yaitu membuat grafik atau plot untuk memvisualisasi data, korelasi antar fitur, dan memahami karakteristik utama untuk setiap variable.
   b. Menganalisis hubungan antar variabel yaitu mencari pola-pola yang penting antar variabel yang mungkin dapat mempengaruhi segmentasi pelanggan.
   c. Identifikasi tren atau pola yaitu mencari anomali, otlier atau perilaku pelanggan yang penting didalam data.
7. Pemodelan Algoritma
   Penelitian ini menggunakan tiga algoritma clustering yaitu: Random Forest, Support Vector Machine (SVM), dan Gradient Boosting. Algoritma ini digunakan untuk melakukan pengelompokkan pelanggan. Ketiga algoritma ini dipilih karena kekuatan masing-masing dalam menghasilkan segmentasi yang akurat dan bermanfaat. Pada tahap ini, model dibangun dengan dilatih menggunakan data training yang telah ada.
8. Evaluasi Model
   Pada penelitian ini, evaluasi model digunakan untuk mengukur kinerja model yang telah dibangun. Dalam evaluasi model digunakan metriks evaluasi seperti akurasi, precision, recall, dan F-Score untuk mengukur seberapa baik model memprediksi segmentasi pelanggan. Selanjutnya dapat dilihat hasil analisis model yang telah dibangun serta memahami kekuatan dan kelemahan model berdasarkan hasil evaluasi.
9. Visualisasi Hasil
   Hasil dari pemodelan yang telah dilakukan ditahap sebelumnya dilakukan visualisai agar dapat memudahkan pengguna dalam mamahami hasil dari klasifikasi yang telah dilakukan [16]. Hasil dari visualisasi model dapat berupa grafik batang, grafik garis, grafik lingkaran, histogram dan lain sebagainya.

IV. HASIL DAN PEMBAHASAN

A. Reading data
   Sebelum melakukan tahapan reading data langkah awal yang harus dilakukan yaitu menginport library yang akan digunakan dalam penelitian ini. Berikut ini merupakan gambar dari tahapan import library :



Gambar 2. Import Library

Gambar 2 merupakan proses dari import library yang dibutuhkan dalam penelitian ini, Library yang telah diimport adalah sebagai berikut :

a. Pandas : digunakan untuk manipulasi dan analisis data.
b. Numpy : digunakan untuk operasi numeric.
c. Seaborn : digunakan untuk visualisasi data.
d. Matplotlib.pyplot : digunakan untuk membuat grafik.

Setelah library diimport, langkah selanjutnya adalah reading data dalam format csv. Pada tahap ini, membaca file CSV yang berisi data pelanggan dan menyimpannya ke dalam DataFrame pandas.

B. Preprocessing data

Tahapan yang dilakukan pada preprocessing data yaitu melakukan eksplorasi struktur data dengan menampilkan data informasi dari data testing yang akan digunakan dalam penelitian ini seperti jumlah atribut yang digunakan beserta type data yang tersedia. Tahap selanjutnya dalam preprocessing data adalah melihat nilai yang kosnog (missing value) dalam tiap-tiap atribut yang akan digunakan dalam penelitian ini. Berikut ini merupakan proses untuk melihat nilai yang kosong. Pada gambar 3 terdapat beberapa atribut yang kosong, yakni: attribute Ever Married, Graduated, Profession, Work_Experience, Family_Size, dan Var_1. Untuk mengatasi *missing*

value pada dataset tersebut, attribut yang kosong diisi dengan jumlah rata-rata dari data yang ada agar dataset yang digunakan sudah bersih dan siap untuk dilakukan analisis.

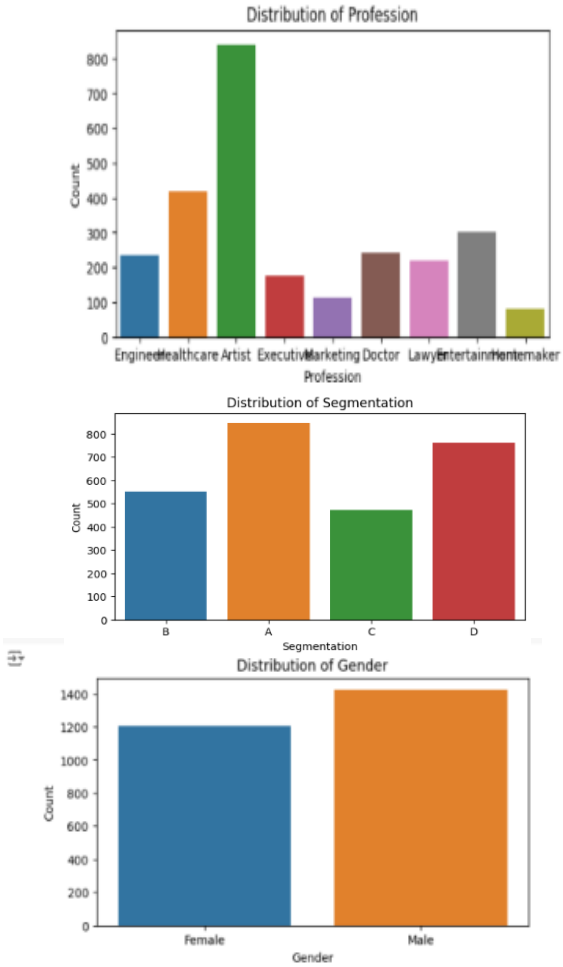
```
missing_values = df_test.isnull().sum()
missing_values
```

```
ID                 0
Gender             0
Ever_Married      50
Age                0
Graduated         24
Profession        38
Work_Experience  269
Spending_Score     0
Family_Size      113
Var_1             32
Segmentation       0
dtype: int64
```

Gambar 3. Proses melihat missing value

C. Exploratori Data Analysis

Hasil explorasi data analysis dalam penelitian ini disajikan pada visualisasi data yang dihasilkan. Visualisasi data ini dapat digunakan untuk menganalisis hubungan antara variabel kategori dengan variabel target yang juga dapat digunakan untuk mengolah pola atau tren yang tersembunyi dalam segmentasi pelanggan. Berikut ini merupakan tampilan visualisasi data yang dihasilkan :



Gambar 4. Visualisasi Data

D. Modeling

Tahap modeling menggunakan tiga algoritma yang berbeda. Berikut ini merupakan algoritma yang digunakan dalam penelitian ini :

1. Model Random Forest

Pada penelitian ini, algoritma Random Forest digunakan untuk memprediksi segmen pelanggan berdasarkan 10 variabel atribut dan 1 menjadi atribut target. Random Forest dipilih karena kemampuannya dalam menangani data dengan dimensi yang tinggi serta mengatasi *overfitting* melalui penggabungan berbagai pohon keputusan. Berikut ini merupakan hasil dari proses modeling menggunakan random forest.

```
=== Model Random Forest ===
Rata-rata Akurasi pada Validasi Silang (k=10): 0.48
Akurasi pada data pengujian: 0.31

Laporan Klasifikasi pada data pengujian:
              precision    recall  f1-score   support

           0       0.34      0.26      0.29       846
           1       0.23      0.24      0.23       550
           2       0.23      0.31      0.26       472
           3       0.40      0.41      0.41       759

    accuracy                           0.31      2627
   macro avg       0.30      0.30      0.30      2627
weighted avg       0.31      0.31      0.31      2627
```

Gambar 5. Matriks Evaluasi Model Random Forest

Hasil dari matrik evaluasi, hasil validasi silang (k=10), model mendapatkan akurasi rata-rata 48%, namun pada data pengujian akurasi menurun menjadi 31%. Perbedaan ini menunjukkan kemungkinan terjadinya overfitting, di mana model bekerja baik pada data pelatihan namun gagal mengeneralisasi data baru. Hasil laporan klasifikasi pada data pengujian menunjukkan bahwa performa model tidak merata di setiap segmen. Precision dan recall untuk segmen 0, 1, dan 2 sangat rendah (di bawah 35%), sedangkan segmen 3 menunjukkan performa yang sedikit lebih baik dengan nilai precision dan recall sekitar 40%. Rata-rata metrik seperti F1-score untuk keseluruhan kelas hanya mencapai 0.30, yang memperkuat bukti bahwa model tidak cukup efektif untuk klasifikasi multikategori ini.

2. Support Vector Machine

Pada penelitian ini, algoritma SVM (Support Vector Machine) digunakan untuk memprediksi segmen pelanggan berdasarkan 10 atribut dan 1 atribut target. SVM dipilih karena kemampuannya dalam menemukan hyperplane optimal yang memisahkan data dalam ruang multidimensi, sehingga cocok untuk menangani masalah klasifikasi yang kompleks. Hasil dari

proses modeling menggunakan SVM disajikan pada Gambar 6.

```
=== Model SVM ===
Rata-rata Akurasi pada Validasi Silang (k=10): 0.45
Akurasi pada data pengujian: 0.35

Laporan Klasifikasi pada data pengujian:
              precision    recall  f1-score   support

           0       0.36      0.35      0.35       846
           1       0.21      0.05      0.09       550
           2       0.26      0.48      0.34       472
           3       0.45      0.48      0.46       759

    accuracy                           0.35      2627
   macro avg       0.32      0.34      0.31      2627
weighted avg       0.34      0.35      0.33      2627
```

Gambar 6. Matriks Evaluasi Model SVM

Hasil dari pengujian Pada model SVM, akurasi rata-rata dari validasi silang (k=10) adalah 45%, sementara akurasi pada data pengujian mencapai 35%. Performa model menunjukkan bahwa kemampuan prediksi SVM tidak optimal, terutama pada data pengujian yang memiliki akurasi rendah. Meskipun akurasi validasi silang mendekati setengah, kemampuan generalisasi model terhadap data baru masih rendah. Dari laporan klasifikasi, terlihat bahwa segmen 0 dan 3 memiliki kinerja yang relatif lebih baik dibandingkan segmen lain, dengan nilai precision dan recall sekitar 35-45%. Segmen 2 menunjukkan recall yang lebih tinggi (48%), namun precision yang lebih rendah (26%), yang berarti model sering salah dalam mengidentifikasi pelanggan di segmen ini. Segmen 1 memiliki performa yang sangat buruk dengan precision 21% dan recall 5%, menunjukkan bahwa model kesulitan mengklasifikasikan pelanggan di segmen ini secara akurat. Secara keseluruhan, hasil evaluasi menunjukkan bahwa model SVM tidak mampu menangani variasi antar segmen dengan baik, terutama untuk segmen yang lebih sulit diidentifikasi seperti segmen 1. Rata-rata f1-score sebesar 31% dan akurasi pengujian yang rendah memperlihatkan bahwa model ini memerlukan optimasi lebih lanjut, baik dalam hal pemilihan fitur maupun tuning parameter untuk meningkatkan performanya.

3. Gradient Boosting

Pada penelitian ini, algoritma Gradient Boosting digunakan untuk memprediksi segmen pelanggan berdasarkan 10 atribut dan 1 atribut target. Gradient Boosting dipilih karena kemampuannya dalam membangun model secara bertahap dengan meminimalkan kesalahan dari model sebelumnya, sehingga menghasilkan prediksi yang lebih akurat melalui pendekatan boosting. Berikut ini merupakan hasil dari proses modeling menggunakan Gradient Boosting.

```
=== Model Gradient Boosting ===
Rata-rata Akurasi pada Validasi Silang (k=10): 0.54
Akurasi pada data pengujian: 0.34

Laporan Klasifikasi pada data pengujian:
              precision    recall  f1-score   support

           0       0.36      0.30      0.32       846
           1       0.26      0.23      0.24       550
           2       0.25      0.33      0.29       472
           3       0.43      0.46      0.44       759

    accuracy                           0.34      2627
   macro avg       0.32      0.33      0.32      2627
weighted avg       0.34      0.34      0.34      2627
```

Gambar 7. Matriks Evaluasi Model Gradient Boosting

Hasil dari model Gradient Boosting, akurasi rata-rata dari validasi silang (k=10) mencapai 54%, yang menunjukkan performa yang lebih baik dibandingkan dengan model lain seperti SVM atau Random Forest pada data pelatihan. Namun, akurasi pada data pengujian hanya sebesar 34%, yang menunjukkan penurunan performa saat diterapkan pada data baru. Hal ini bisa mengindikasikan bahwa model mengalami overfitting, di mana model bekerja baik pada data pelatihan, tetapi kurang efektif dalam menggeneralisasi data pengujian. Berdasarkan laporan klasifikasi, segmen 0 dan 3 memiliki kinerja yang lebih baik dibandingkan segmen lainnya. Precision dan recall untuk segmen 3 mencapai masing-masing 43% dan 46%, yang merupakan kinerja terbaik di antara keempat segmen. Di sisi lain, segmen 1 dan 2 memiliki precision dan recall yang relatif rendah, di mana segmen 1 hanya memiliki precision 26% dan recall 23%, menunjukkan bahwa model kesulitan mengklasifikasikan pelanggan di segmen ini. Secara keseluruhan, performa model Gradient Boosting masih kurang memuaskan dengan rata-rata f1-score sekitar 32%. Walaupun model menunjukkan potensi dengan akurasi validasi silang yang lebih tinggi dibanding model lain, performa yang rendah pada data pengujian menunjukkan bahwa optimasi lebih lanjut diperlukan, baik dalam hal tuning hyperparameter atau penanganan ketidakseimbangan kelas agar dapat memberikan prediksi yang lebih akurat di semua segmen pelanggan.

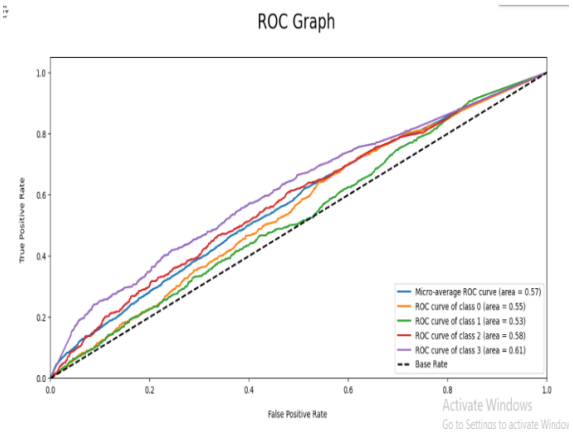
E. Evaluasi Model

Berdasarkan hasil pemodelan yang telah dilakukan ditahap sebelumnya, hasil dari rata-rata

akurasi pada validasi silang menggunakan K=10 menghasilkan nilai yang cukup beragam. Dengan nilai yang dihasilkan oleh Random Forest sebesar 0,48, sementara untuk nilai yang dihasilkan oleh model Support Vector Machine untuk nilai rata-rata yang dihasilkan sebesar 0,45. Dan untuk model Gradient Boosting rata-rata akurasi yang dihasilkan pada validasi silang sebesar 0,54. Setalah melihat hasil dari masing-masing model dapat kita simpulkan bahwa model yang memiliki nilai rata-rata validasi yang cukup tinggi yaitu model Gradient Boosting dengan nilai 0.54. sementara untuk nilai dari kedua model yang lain menghasilkan nilai yang hanya memiliki selisih sedikit yaitu antara 0,48 untuk model Random Forest serta 0,45 untuk model Support Vector Mechine.

F. Visualisasi Model

Visualisasi model yang dihasilkan dalam penelitian ini yaitu menggunakan ROC AUC agar dapat membandingan nilai tiap=tiap segmentasi yang dihasilkan. Berikut ini merupakan tampilan yang dihasilkan dengan ROC AUC.

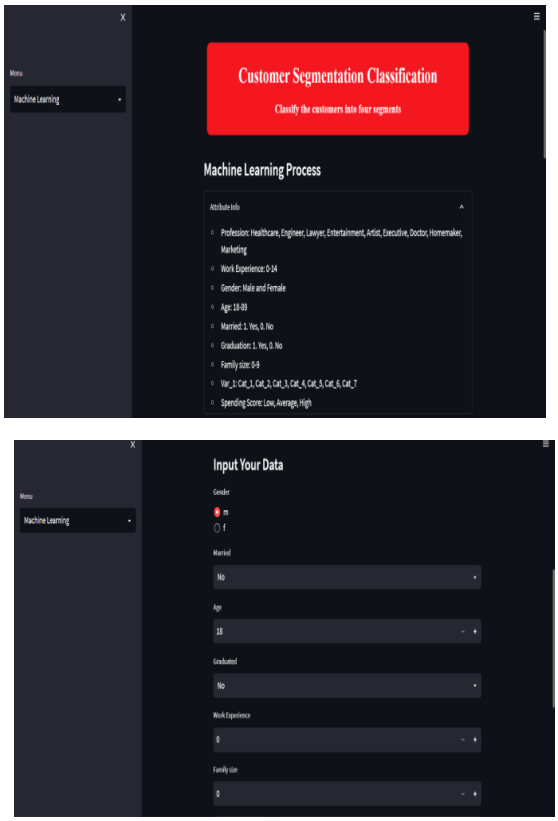


Gambar 8. Visualisasi ROC Graph

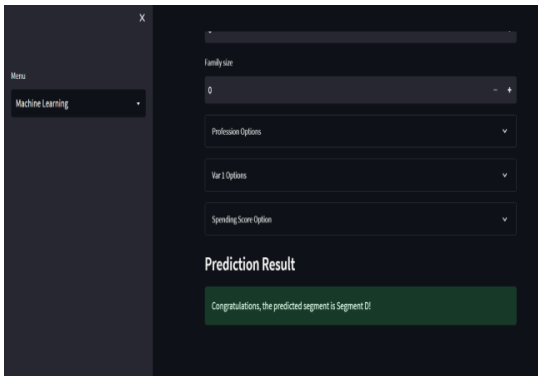
Setelah melihat tampilan dari visualisi model yang dihasilkan, dapat dilihat juga skor yang dihasilkan dalam tiap-tiap sementasi. Pada Gambar 8 menampilkan skor yang dihasilkan dalam tiap-tiap segmentasi. Setelah melihat skronya dapat kita lihat bahwa skor tertinggi dihaslikan oleh segmentasi D dengan nilai 0,61. Segmentasi D memiliki skor yang paling tinggi diantara ketiganya, sementara itu skor terendah dihasilkan oleh segmentasi B dengan skor 0,52.

G. Deployment Model

Dalam penelitian ini, Streamlit digunakan sebagai platform untuk mendeploy model segmentasi pelanggan yang telah dikembangkan sebelumnya. Dengan menggunakan *Streamlit,* model dapat diakses melalui antarmuka web yang interaktif, memungkinkan pengguna untuk melakukan segemntasi pelanggan berdasarkan input variabel yang disediakan. Pendekatan ini memudahkan dalam memvisualisasikan hasil clustering dan mempercepat proses pengujian serta validasi model di lingkungan yang lebih praktis dan *user-friendly*. Berikut merupakan tampilan dari deploy model:



Gambar 9. Tampilan pada menu "Machine Learning"



Gambar 10. Tampilan pada menu "Result"

Tampilan menu "Machine Learning" untuk melakukan Input data, seperti: Gender, Married, Age, Graduated, Work experience, Family Size,

profesi, var 1, dan spending skor. Setelah semua permintaan inputan telah terisi, akan di tampilan hasil prediksi segmentasi customer seperti pada gambar 10.

V. KESIMPULAN

Penelitian ini berhasil menerapkan tiga algoritma pembelajaran mesin, yaitu *Random Forest, Support Vector Machine (SVM)*, dan *Gradient Boosting,* untuk memprediksi segmen pelanggan berdasarkan 10 atribut. Hasil menunjukkan bahwa *Gradient Boosting* memiliki akurasi terbaik pada validasi silang dengan rata-rata 54%, sedangkan *Random Forest* dan SVM masing-masing mencapai 48% dan 45%. Namun, semua model mengalami penurunan akurasi pada data pengujian, dengan *Gradient Boosting* tetap unggul meski hanya mencapai akurasi 34%.

Meskipun *Gradient Boosting* menunjukkan performa yang lebih baik dibandingkan model lain, hasil pengujian keseluruhan masih kurang memuaskan, terutama karena adanya indikasi *overfitting* dan ketidakseimbangan dalam performa klasifikasi di tiap segmen. Oleh karena itu, diperlukan optimasi lebih lanjut, seperti tuning hyperparameter dan penanganan data yang lebih baik, untuk meningkatkan kemampuan generalisasi model.

REFERENSI

[1] K. Abdullah *et al.*, *Metodologi Penelitian Kuantitatif Metodologi Penelitian Kuantitatif*. Aceh, 2021.

[2] I. Afdhal, R. Kurniawan, I. Iskandar, R. Salambue, E. Budianita, and F. Syafria, "Penerapan Algoritma Random Forest Untuk Analisis Sentimen Komentar Di YouTube Tentang Islamofobia," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 5, no. 1, pp. 122–130, 2022, [Online]. Available: http://ojs.serambimekkah.ac.id/jnkti/article/view/4004/pdf

[3] N. H. Harani, C. Prianto, and F. A. Nugraha, "Segmentasi Pelanggan Produk Digital Service Indihome Menggunakan Algoritma K-Means Berbasis Python," *J. Manaj. Inform.*, vol. 10, no. 2, pp. 133–146, 2020, doi: 10.34010/jamika.v10i2.2683.

[4] K. Kraugusteeliana, S. Muis, F. Nugroho, A. Karim, and Y. Siagian, "Data Mining Klasifikasi Breast Cancer Menerapkan Algoritma Gradient Boosted Trees," *J. MEDIA Inform. BUDIDARMA Vol. 7, Nomor 2, April 2023, Page 881-890*, vol. 7, no. April, pp. 881–890, 2023, doi: 10.30865/mib.v7i2.6095.

[5] I. Kurniawan, D. C. P. Buani, A. Abdussomad, W. Apriliah, and R. A. Saputra, "Implementasi Algoritma Random Forest Untuk Menentukan Penerima Bantuan Raskin," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 2, pp. 421–428, 2023, doi: 10.25126/jtiik.20231026225.

[6] F. R. Lumbanraja, B. S. Ira Hariati, D. Kurniawan, and A. Aristoteles, "Prediksi Jumlah Penderita Penyakit Tuberkulosis Di Kota Bandar Lampung Menggunakan Metode Svm (Support Vector Machine)," *Kumpul. J. Ilmu Komput.*, vol. 7, no. 3, pp. 320–330, 2020.

[7] D. P. Utomo and M. Mesran, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," *J. Media Inform. Budidarma*, vol. 4, no. 2, p. 437, 2020, doi: 10.30865/mib.v4i2.2080.

[8] M. I. Fianty, M. E. Johan, A. Aulia, and M. M. Veronica, "Application of Clustering-Based Data Mining for the Assessment of Nutritional Status in Toddlers at Community Health Centers," *J. Inf. Syst. Informatics*, vol. 5, no. 4, pp. 1350–1362, Dec. 2023, doi: 10.51519/journalisi.v5i4.586.

[9] A. C. Puspita, B. Setiawan, C. Ayu, J. Heikal, and U. Bakrie, "Analisis Preferensi Konsumen dan Profil Pembeli di Industri Otomotif : Studi Kasus Pembelian Kendaraan di Wilayah Jabodetabek," *J. Mhs. Ekon. Bisnis*, vol. 4, no. 3, pp. 1036–1050, 2024.

[10] A. Faadhilah and H. Nugroho, "Pemetaan Daerah Rawan Longsor di Kabupaten Bandung Barat menggunakan Metode Machine Learning dengan Teknik SVM," *Rekayasa Hijau J. Teknol. Ramah Lingkung.*, vol. 8, no. 2, 2024.

[11] M. R. Givari, M. R. Sulaeman, and Y. Umaidah, "Perbandingan Algoritma SVM, Random Forest Dan XGBoost Untuk Penentuan Persetujuan Pengajuan Kredit," *Nuansa Inform.*, vol. 16, no. 1, pp. 141–149, 2022, doi: 10.25134/nuansa.v16i1.5406.

[12] M. Riyyasy, Azfa, Rasikh, W. Aghniya, Nouval, and H. Tantyoko, "Penerapan Algoritma Machine Learning Untuk Memprediksi Term Deposit Nasabah Perbankan," *J. Inform. Inf. Technol.*, vol. 2, no. 3, pp. 145–156, 2023, [Online]. Available: https://journal.ittelkom-pwt.ac.id/index.php/ledger/article/view/1416

[13] A. Primajaya and B. N. Sari, "Random forest algorithm for prediction of precipitation," *Indones. J. Artif. Intell. Data Min.*, vol. 1, no. 1, pp. 27–31, 2018.

[14] S. Aryana, "Studi Literatur: Analisis Penerapan dan Pengembangan Penilaian Autentik Kurikulum 2013 pada Jurnal Nasional dan Internasional," *Pros. Semin. Nas. Pascasarj.*, vol. 4, no. 1, pp. 368–374, 2021.

[15] G. S. G. Prawira and H. Setiaji, "Penerapan Data Transformation Pada Database Sistem Informasi Manajemen Rumah Sakit," *Sintak*, vol. 3, no. 0 SE-Vol 3 (2019), pp. 1–5, 2019, [Online]. Available: https://unisbank.ac.id/ojs/index.php/sintak/article/view/7595

[16] M. D. Muafa, "Pengembangan Aplikasi Berbasis Web dengan Rshiny untuk Data Klasifikasi Menggunakan Metode Naive Bayes," *Automata*, vol. 3, no. 1, p. 8, 2022, [Online]. Available: https://journal.uii.ac.id/AUTOMATA/article/view/21875