

Perbandingan Algoritma *Random Forest*, *Logistic Regression*, *Support Vector Machine*, dan *Gradient Boosting* untuk Klasifikasi Popularitas Lagu Menggunakan TF-IDF

Fikko Nanda Ramadhani¹, Naaq Shal Ramadanu^{2*}, Adit Ramadan³,
Muhammad Ridwan Martana Putra⁴, Aprilisa Arum Sari⁵

¹Teknik Informatika
Universitas Duta Bangsa
¹fikko.fitto16@gmail.com

²Teknik Informatika
Universitas Duta Bangsa
^{2*}naaqdanu@gmail.com

³Teknik Informatika
Universitas Duta Bangsa
³aditmdn117@gmail.com

⁴Teknik Informatika
Universitas Duta Bangsa
⁴ridwanmartana5@gmail.com

⁵Teknik Informatika
Universitas Duta Bangsa
⁵aprilisa_arumsari@udb.ac.id

Abstrak— Perkembangan platform musik digital menghasilkan data dalam jumlah besar yang dapat dimanfaatkan untuk memperoleh informasi melalui teknik data mining, salah satunya klasifikasi. Penelitian ini bertujuan membandingkan performa algoritma *Random Forest*, *Logistic Regression*, *Support Vector Machine (SVM)*, dan *Gradient Boosting* dalam mengklasifikasikan popularitas lagu menggunakan dataset Top 500 Songs. Tahapan penelitian meliputi preprocessing data, pembentukan label berdasarkan posisi lagu pada tangga lagu menjadi kategori Top 10 dan Non Top 10, transformasi data menggunakan TF-IDF pada atribut deskripsi lagu, standardisasi fitur numerik menggunakan *StandardScaler*, pembagian data dengan rasio 80% data latih dan 20% data uji, serta pembangunan model klasifikasi. Evaluasi dilakukan menggunakan matrix *Accuracy*, *Precision*, *Recall*, *F1-Score*, dan *Confusion Matrix*. Hasil pengujian menunjukkan bahwa *Logistic Regression* memperoleh performa terbaik dengan nilai akurasi sebesar 84%, diikuti *Support Vector Machine* sebesar 82%, *Random Forest* sebesar 80%, dan *Gradient Boosting* sebesar 78%. Hasil penelitian menunjukkan bahwa algoritma *Logistic Regression* lebih efektif dibandingkan algoritma lainnya dalam mengklasifikasikan popularitas lagu pada dataset yang direpresentasikan menggunakan TF-IDF. Penelitian ini diharapkan dapat menjadi referensi dalam pemilihan algoritma klasifikasi untuk data teks pada bidang musik digital.

Kata kunci— Klasifikasi Teks, TF-IDF, *Logistic Regression*, Machine Learning, Popularitas Lagu .

Abstract— The growth of digital music platforms generates vast amounts of data that can be leveraged through data mining techniques such as classification to extract valuable insights. This study aims to compare the performance of *Random Forest*, *Logistic Regression*, *Support Vector Machine (SVM)*, and *Gradient Boosting* algorithms in classifying song popularity using the "Top 500 Songs" dataset. The research process involved data preprocessing; label creation based on chart rankings (categorized as "Top 10" and "Non-Top 10"); data transformation using TF-IDF on song description attributes; numerical feature standardization via *StandardScaler*; data splitting (80% training, 20% testing); and classification model development. Performance was evaluated using *Accuracy*, *Precision*, *Recall*, *F1-Score*, and *Confusion Matrix* metrics. The results indicate that *Logistic Regression* achieved the best performance with an *Accuracy* of 84%, followed by *Support Vector Machine* (82%), *Random Forest* (80%), and *Gradient Boosting* (78%). These findings demonstrate that *Logistic Regression* is more effective than the other algorithms for classifying song popularity in datasets represented using TF-IDF. This study is expected to serve as a reference for selecting classification algorithms for text-based data in the digital music domain.

Keywords— Text Classification, TF-IDF, *Logistic Regression*, Machine Learning, Song Popularity.

I. PENDAHULUAN

Perkembangan platform musik digital telah menghasilkan data dalam jumlah besar yang terus bertambah setiap hari. Data tersebut tidak hanya berupa data numerik, tetapi juga data tekstual seperti deskripsi lagu yang dapat dimanfaatkan untuk memperoleh informasi dan pengetahuan baru melalui proses data mining. Salah satu

teknik data mining yang banyak digunakan adalah klasifikasi, yaitu proses mengelompokkan data ke dalam kategori tertentu berdasarkan karakteristik yang dimiliki data[1].

Pada industri musik, popularitas lagu menjadi salah satu indikator penting untuk mengukur keberhasilan suatu karya. Popularitas tersebut umumnya ditunjukkan melalui posisi pada tangga lagu, lama bertahan pada chart, maupun

tingkat penerimaan pendengar. Analisis terhadap data historis lagu dapat digunakan untuk mengidentifikasi pola-pola yang berkaitan dengan popularitas lagu sehingga dapat dimanfaatkan sebagai pendukung pengambilan keputusan berbasis data [2].

Klasifikasi data teks telah berkembang pesat dan banyak diterapkan pada berbagai bidang, seperti analisis sentimen, klasifikasi berita, deteksi spam, dan sistem rekomendasi. Penelitian sebelumnya menunjukkan bahwa representasi teks menggunakan Term Frequency-Inverse Document Frequency (TF-IDF) masih memberikan performa yang kompetitif, terutama ketika dikombinasikan dengan algoritma machine learning klasik seperti *Logistic Regression*, *Support Vector Machine (SVM)*, dan *Random Forest* [3], [4].

Berbagai algoritma klasifikasi memiliki karakteristik dan tingkat performa yang berbeda pada data tekstual [5]. *Logistic Regression* dikenal memiliki proses pelatihan yang sederhana dan mudah diinterpretasikan, *SVM* memiliki kemampuan yang baik dalam menangani data berdimensi tinggi, *Random Forest* mampu mengurangi risiko overfitting melalui pembentukan banyak pohon keputusan, sedangkan *Gradient Boosting* membangun model secara bertahap untuk meningkatkan akurasi prediksi [6]. Oleh karena itu, diperlukan analisis komparatif untuk mengetahui algoritma yang paling efektif dalam menyelesaikan permasalahan klasifikasi pada data popularitas lagu.

Meskipun berbagai penelitian telah menerapkan TF-IDF pada klasifikasi teks, sebagian besar penelitian hanya membandingkan dua atau tiga algoritma klasifikasi. Selain itu, penelitian mengenai klasifikasi popularitas lagu menggunakan kombinasi TF-IDF dengan empat algoritma machine learning klasik masih relatif terbatas. Oleh karena itu penelitian ini melakukan analisis komprehensif terhadap empat algoritma tersebut.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan membandingkan performa empat algoritma klasifikasi, yaitu *Random Forest*, *Logistic Regression*, *Support Vector Machine*

(*SVM*), dan *Gradient Boosting* dalam mengklasifikasikan popularitas lagu menggunakan dataset Top 500 Songs. Evaluasi model dilakukan menggunakan metrik *Accuracy*, *Precision*, *Recall*, *F1-Score*, dan Confusion Matrix. Hasil penelitian diharapkan dapat memberikan gambaran mengenai algoritma yang memiliki performa terbaik untuk diterapkan pada kasus klasifikasi popularitas lagu berbasis data historis

II. METODOLOGI PENELITIAN

Metodologi penelitian merupakan tahapan yang digunakan untuk mencapai tujuan penelitian secara sistematis. Pada penelitian ini diterapkan pendekatan data mining dengan teknik klasifikasi untuk mengidentifikasi popularitas lagu berdasarkan karakteristik yang terdapat pada dataset. Tahapan penelitian secara keseluruhan dimulai dari pengumpulan dataset Top 500 Songs, dilanjutkan dengan proses preprocessing data, transformasi data menggunakan TF-IDF dan StandardScaler, pembagian dataset menjadi data latih (80%) dan data uji (20%), penerapan empat algoritma klasifikasi, evaluasi model menggunakan beberapa metrik pengukuran, hingga penentuan algoritma terbaik berdasarkan hasil pengujian.

Untuk memastikan tahapan berjalan secara terstruktur, penelitian ini menggunakan Metode CRISP-DM (Cross-Industry Standard Process for Data Mining). Metode ini digunakan sebagai kerangka kerja utama yang mencakup beberapa fase standar, mulai dari pemahaman terhadap data (data understanding), pembersihan dan penyiapan data (data preparation), pengembangan model klasifikasi (modeling), hingga pengujian performa algoritma (evaluation). Tahapan penelitian secara rinci berdasarkan kerangka kerja ini dijelaskan pada subbab berikut.

A. Jenis Penelitian

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan Data Mining menggunakan teknik klasifikasi untuk membangun model prediksi berdasarkan data historis yang telah diketahui kelasnya [7]. Tujuan penelitian adalah membandingkan performa empat

algoritma klasifikasi, yaitu *Random Forest*, *Logistic Regression*, *Support Vector Machine (SVM)*, dan *Gradient Boosting* dalam mengelompokkan lagu ke dalam kategori Top 10 dan Non Top 10 berdasarkan karakteristik yang dimiliki.

Metode klasifikasi dipilih karena mampu membangun model prediksi berdasarkan data historis yang telah diketahui kelasnya. Hasil penelitian diharapkan dapat menunjukkan algoritma yang memiliki performa terbaik dalam mengidentifikasi popularitas lagu.

B. Sumber dan Pengumpulan Data

Dataset yang digunakan dalam penelitian ini adalah *Top 500 Songs Dataset* yang diperoleh dari Kaggle. Data yang digunakan merupakan data sekunder berupa dataset publik dan digunakan sebagai objek penelitian untuk proses klasifikasi popularitas lagu. Dataset tersebut memuat beberapa atribut penting seperti judul lagu (*title*), deskripsi lagu (*description*), nama album (*appears on*), nama artis (*artist*), penulis lagu (*writers*), produser (*producer*), waktu rilis (*released*), lama bertahan pada chart (*streak*), dan posisi pada tangga lagu (*position*) [8]. Contoh sampel data mentah sebelum dilakukan proses *preprocessing* dapat dilihat pada Tabel 1.

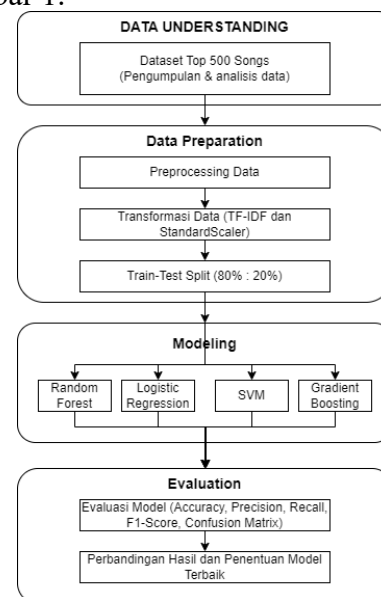
Tabel 1. Sampel Data Mentah *Top 500 Songs Dataset*

title	description	Appears on	artist	writers	producer	released	streak	position
Like a Rolling Stone	"I wrote it. I didn't fail. It was straightforward."	Highway 61 Revisited (Columbia)	Bob Dylan	Dylan	Tom Wilson	July 1965	12 weeks	No. 2
(I Can't Get No) Satisfaction	"It's the riff I heard around the world," says St	Out of Our Heads (ABKCO)	The Rolling Stones	Mick Jagger, Keith Richards	Andrew Loog Oldham	May 1965	14 weeks	No. 1
Imagine	John Lennon wrote	Imagine (Capitol/Apple)	John Lennon	John Lennon	Lennon, Phil Spector	October 1971	9 weeks	No. 3

	"Imagine, his greatest music"				actor, Yo			
					ko On o			
What's Going On	"What's Going On" is an exquisite plea for peace	What's Going On (Tamla)	Marvin Gaye	Gaye, Renaldo Benson, Al Cleveland	Ga ye	February 1971	13 weeks	No. 2
Respect	Otis Redding wrote "Respect" and recorded it f	I Never Loved a Man the Way I Love You (Atlantic)	Aretha Franklin	Otis Redding	Jerry Wexler	April 1967	12 weeks	No. 1

C. Tahap Penelitian

Penelitian ini dilakukan melalui beberapa tahapan yang dimulai dari pengumpulan dataset Top 500 Songs, preprocessing data, transformasi data menggunakan TF-IDF dan StandardScaler, pembagian data menjadi data latih dan data uji, penerapan empat algoritma klasifikasi, evaluasi model, hingga penentuan algoritma terbaik. Alur penelitian secara keseluruhan ditunjukkan pada Gambar 1.



Gambar 1. Alur penelitian klasifikasi popularitas lagu menggunakan TF-IDF dan empat algoritma machine learning.

D. Preprocessing Data

Tahap *preprocessing* dilakukan untuk meningkatkan kualitas data sebelum diproses oleh algoritma *machine learning* [9]. Data yang tidak bersih atau tidak lengkap dapat memengaruhi hasil klasifikasi, sehingga diperlukan proses pembersihan data terlebih dahulu. Hasil dari tahapan *preprocessing* data ini menghasilkan struktur data baru yang siap digunakan untuk pemodelan, sebagaimana ditunjukkan pada Tabel 2.

Tabel 2. Sampel Atribut Hasil *Preprocessing* Data

description	streak	weeks	year	is top 10
"I wrote it. I didn't fail. It was straight,"	12		1965	1
"It's the riff heard round the world," says St.	14		1965	1
John Lennon wrote "Imagine," his greatest mus	9		1971	1
"What's Going On" is an exquisite plea for pea	13		1971	1
Otis Redding wrote "Respect" and recorded it f	12		1967	1

Adapun tahapan processing yang dilakukan meliputi:

1. Pembuatan Label Klasifikasi

Atribut *position* digunakan sebagai dasar pembentukan kelas target (*target label*). Lagu yang memiliki posisi 1 sampai 10 dikategorikan sebagai Top 10 dan diberi label 1, sedangkan lagu dengan posisi di atas 10 dikategorikan sebagai Non Top 10 dan diberi label 0.

Proses ini dilakukan untuk mengubah data posisi chart menjadi bentuk klasifikasi biner yang dapat diproses oleh algoritma machine learning.

2. Transformasi Data Numerik

Atribut *streak* yang menunjukkan lama lagu bertahan pada chart diubah menjadi bentuk numerik berupa jumlah minggu (*streak_weeks*). Transformasi ini bertujuan agar data dapat digunakan sebagai fitur dalam proses klasifikasi.

3. Ekstraksi Tahun Rilis

Atribut *released* diproses untuk mengambil informasi tahun rilis lagu. Informasi tahun digunakan sebagai salah

satu variabel yang diduga memiliki pengaruh terhadap popularitas lagu.

4. Penanganan Missing Value

Data yang memiliki nilai kosong pada atribut *description* dihapus karena atribut tersebut merupakan fitur utama dalam penelitian. Sedangkan nilai kosong pada atribut tahun rilis diisi menggunakan nilai median agar tidak mengurangi jumlah data yang tersedia.

E. Transformasi dan Representasi Data

Dataset yang digunakan terdiri dari data teks dan data numerik sehingga diperlukan proses transformasi sebelum dilakukan klasifikasi.

1. TF-IDF (Term Frequency-Inverse Document Frequency)

Atribut *description* berupa teks diubah menjadi representasi numerik menggunakan metode TF-IDF [10]. Metode ini memberikan bobot pada setiap kata berdasarkan frekuensi kemunculannya dalam dokumen dan tingkat keunikannya pada keseluruhan dataset.

Penggunaan TF-IDF memungkinkan informasi penting yang terkandung dalam deskripsi lagu dapat direpresentasikan dalam bentuk angka sehingga dapat diproses oleh algoritma machine learning.

2. Standardisasi Data

Atribut numerik berupa *streak_weeks* dan *year* dinormalisasi menggunakan Standard Scaler Agar setiap fitur memiliki skala yang sebanding sebelum proses pelatihan model [11]. Proses ini bertujuan untuk menyamakan skala antar fitur sehingga tidak terjadi dominasi nilai pada atribut tertentu saat proses pelatihan model.

F. Pembagian Dataset

Dataset yang telah diproses kemudian dibagi menjadi dua bagian menggunakan metode Train-Test Split [11].

1. 80% data digunakan sebagai data latih (*training data*)

2. 20% data digunakan sebagai data uji (*testing data*)

Data latih digunakan untuk membangun model klasifikasi, sedangkan data uji digunakan untuk mengukur kemampuan model dalam melakukan prediksi terhadap data yang belum pernah dipelajari sebelumnya.

G. Penerapan Algoritma Klasifikasi

Pada penelitian ini dilakukan perbandingan empat algoritma klasifikasi yang memiliki karakteristik berbeda.

1. *Random Forest*

Random Forest merupakan algoritma berbasis *ensemble learning* yang membangun sejumlah pohon keputusan (*decision tree*) dan menggabungkan hasil prediksi dari seluruh pohon untuk menghasilkan keputusan akhir. Algoritma ini dipilih karena memiliki kemampuan yang baik dalam menangani data berdimensi tinggi dan mampu mengurangi risiko overfitting.

2. *Logistic Regression*

Logistic Regression merupakan algoritma klasifikasi yang digunakan untuk menghitung probabilitas suatu data termasuk ke dalam kelas tertentu [12]. Algoritma ini sering digunakan sebagai model dasar karena memiliki proses perhitungan yang sederhana dan mudah diinterpretasikan.

3. *Support Vector Machine (SVM)*

Support Vector Machine bekerja dengan mencari garis pemisah (*hyperplane*) terbaik yang mampu memisahkan dua kelas data secara optimal. Algoritma ini banyak digunakan pada kasus klasifikasi teks karena mampu bekerja dengan baik pada data berdimensi tinggi hasil transformasi TF-IDF.

4. *Gradient Boosting*

Gradient Boosting merupakan algoritma *ensemble learning* yang membangun model secara bertahap dengan memperbaiki kesalahan prediksi dari model sebelumnya. Pendekatan ini memungkinkan model menghasilkan performa yang lebih baik dibandingkan metode klasifikasi tunggal.

H. Evaluasi Model

Kinerja setiap algoritma dievaluasi menggunakan beberapa metrik pengukuran.

1. *Accuracy*

Accuracy digunakan untuk mengukur persentase prediksi yang benar dibandingkan dengan keseluruhan data uji.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

2. *Precision*

Precision digunakan untuk mengukur ketepatan model dalam memprediksi kelas positif.

$$Precision = TP / (TP + FP)$$

3. *Recall*

Recall digunakan untuk mengukur kemampuan model dalam menemukan seluruh data yang termasuk ke dalam kelas positif.

$$Recall = TP / (TP + FN)$$

4. *F1-Score*

F1-Score digunakan untuk mengukur keseimbangan antara *Precision* dan *Recall*.

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall)$$

5. *Confusion Matrix*

Confusion Matrix digunakan untuk menampilkan jumlah prediksi benar dan salah pada masing-masing kelas yang lebih detail mengenai performa model.

III. HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil pengujian dan analisis terhadap empat algoritma klasifikasi yang digunakan dalam penelitian, yaitu *Random Forest*, *Logistic Regression*, *Support Vector Machine (SVM)*, dan *Gradient Boosting*. Pengujian dilakukan menggunakan dataset Top 500 Songs yang telah melalui tahap preprocessing dan transformasi data. Proses transformasi dilakukan dengan menerapkan metode TF-IDF pada atribut deskripsi lagu serta standarisasi pada atribut numerik sehingga data dapat diproses secara optimal oleh algoritma machine learning.

Evaluasi model dilakukan menggunakan data uji yang diperoleh dari proses pembagian dataset dengan rasio 80% data latih dan 20% data uji. Kinerja masing-masing algoritma dianalisis

berdasarkan nilai *Accuracy*, *Precision*, *Recall*, dan *F1-Score*. Selain itu, *Confusion Matrix* digunakan untuk memberikan gambaran yang lebih rinci mengenai kemampuan model dalam mengklasifikasikan lagu ke dalam kategori Top 10 dan Non Top 10. Hasil pengujian setiap algoritma disajikan pada subbab berikut.

A. Hasil Pengujian *Random Forest*

Pada tahap ini dilakukan pengujian menggunakan algoritma *Random Forest*. Model dibangun menggunakan fitur hasil transformasi TF-IDF pada atribut *description* serta fitur numerik *streak_weeks* dan *year*.

Berdasarkan hasil pengujian diperoleh nilai akurasi sebesar 80,00%.

```
=====
MODEL: Random Forest
Akurasi: 80.00%
=====
Laporan Klasifikasi Detail:
      precision    recall  f1-score   support

   0       0.87       0.69       0.77        48
   1       0.76       0.90       0.82        52

 accuracy          0.80          100
  macro avg       0.81          100
 weighted avg     0.81          100
=====
```

Gambar 2. Algoritma *Random Forest*

B. Hasil Pengujian *Logistic Regression*

Pengujian berikutnya dilakukan menggunakan algoritma *Logistic Regression*. Algoritma ini digunakan sebagai model klasifikasi linear untuk memprediksi kategori popularitas lagu. Berdasarkan hasil pengujian diperoleh nilai akurasi sebesar 84.00%.

```
=====
MODEL: Logistic Regression
Akurasi: 84.00%
=====
Laporan Klasifikasi Detail:
      precision    recall  f1-score   support

   0       0.88       0.77       0.82        48
   1       0.81       0.90       0.85        52

 accuracy          0.84          100
  macro avg       0.85          100
 weighted avg     0.84          100
=====
```

Gambar 3. *Logistic Regression*

Berdasarkan Gambar 3, Model mampu melakukan klasifikasi dengan cukup meskipun masih terdapat beberapa kesalahan prediksi pada kelas Top 10 maupun Non Top 10.

C. Hasil Pengujian *Support Vektor Machine (SVM)*

Algoritma *SVM* digunakan untuk mencari hyperplane terbaik yang memisahkan kedua kelas data. Berdasarkan hasil pengujian diperoleh nilai akurasi 82.00%

```
=====
MODEL: Support Vector Machine (SVM)
Akurasi: 82.00%
=====
Laporan Klasifikasi Detail:
      precision    recall  f1-score   support

   0       0.86       0.75       0.80        48
   1       0.79       0.88       0.84        52

 accuracy          0.82          100
  macro avg       0.83          100
 weighted avg     0.82          100
=====
```

Gambar 4. *Support Vector Machine (SVM)*

Berdasarkan Gambar 4, model menunjukkan kemampuan yang baik dalam membedakan lagu kategori Top 10 dan Non Top 10. Hal ini terlihat dari dominasi nilai prediksi yang benar dibandingkan prediksi yang salah.

D. Hasil Pengujian *Gradient Boosting*

Pengujian terakhir dilakukan menggunakan algoritma *Gradient Boosting* yang merupakan metode ensemble berbasis boosting.

Berdasarkan hasil pengujian diperoleh nilai akurasi sebesar 78.00%.

```
=====
MODEL: Gradient Boosting
Akurasi: 78.00%
=====
Laporan Klasifikasi Detail:
      precision    recall  f1-score   support

     0       0.86      0.65      0.74        48
     1       0.73      0.90      0.81        52

 accuracy          0.78        100
  macro avg       0.80      0.77      0.77        100
 weighted avg     0.80      0.78      0.78        100
=====
```

Gambar 5. Algoritma *Gradient Boosting*

Berdasarkan gambar 5, model mampu meningkatkan performa klasifikasi melalui proses pembelajaran bertahap yang memperbaiki kesalahan model sebelumnya.

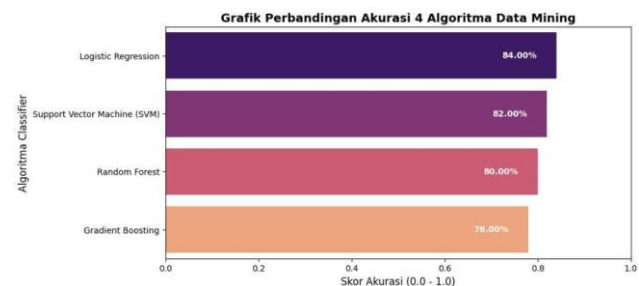
E. Perbandingan Performa Algoritma Setelah seluruh algoritma diuji,

Tabel 3. Perbandingan Performa Algoritma Klasifikasi

Algoritma	Accuracy	Keterangan
<i>Random Forest</i>	80,00%	Mampu menghasilkan performa yang baik, namun masih di bawah <i>Logistic Regression</i> dan <i>SVM</i> .
<i>Logistic Regression</i>	84,00%	Memperoleh akurasi tertinggi sehingga menjadi model terbaik pada penelitian ini.
<i>Support Vector Machine (SVM)</i>	82,00%	Memiliki performa yang baik dan berada pada urutan kedua setelah <i>Logistic Regression</i> .
<i>Gradient Boosting</i>	78,00%	Menghasilkan akurasi paling rendah dibandingkan algoritma lainnya pada dataset ini.

Berdasarkan Tabel 3, algoritma *Logistic Regression* memperoleh nilai akurasi tertinggi sebesar 84,00%, sehingga dipilih sebagai model terbaik dalam penelitian ini. Algoritma *Support Vector Machine (SVM)* berada pada urutan kedua dengan akurasi 82,00%, diikuti *Random Forest* sebesar 80,00%, sedangkan *Gradient Boosting* memperoleh akurasi terendah yaitu 78,00%. Hasil ini menunjukkan bahwa *Logistic Regression* lebih efektif dalam

mengklasifikasikan popularitas lagu berdasarkan fitur yang telah ditransformasikan menggunakan TF-IDF dan StandardScaler. Dilakukan Perbandingan Berdasarkan nilai akurasi yang di peroleh



Gambar 6. Grafik Perbandingan Akurasi Empat Algoritma Klasifikasi

Berdasarkan gambar 6 dapat diketahui bahwa algoritma *Logistic Regression* memperoleh performa terbaik dengan nilai akurasi sebesar 84,00%. Sementara itu algoritma *Gradient Boosting* memperoleh akurasi terendah sebesar 78,00%.

F. Analisis Performa Model Terbaik

Keunggulan Keunggulan *Logistic Regression* dalam mencapai akurasi tertinggi (84,00%) dibandingkan tiga algoritma kompleks lainnya sangat dipengaruhi oleh karakteristik dataset yang digunakan. Atribut description yang ditransformasikan menggunakan metode TF-IDF menghasilkan matriks fitur berdimensi sangat tinggi (high-dimensional data) dan renggang (sparse matrix), di mana jumlah kata unik yang menjadi fitur baru jauh lebih banyak daripada jumlah sampel datanya.

Pada kondisi data seperti ini, hubungan antarfitur teks cenderung bersifat linear terhadap label popularitas (*is_top_10*). *Logistic Regression* bekerja sangat efisien pada ruang berdimensi tinggi karena secara inheren melakukan pemisahan linear dan meminimalkan risiko *overfitting* melalui fungsi regularisasi yang sederhana.

Di sisi lain, algoritma berbasis pohon seperti *Random Forest* (80,00%) dan *Gradient Boosting* (78,00%) cenderung kurang optimal dan rentan mengalami penurunan performa pada data renggang (sparse data) hasil TF-IDF. Struktur

percabangan pohon keputusan (*decision tree*) kesulitan menentukan batas potong (*split*) yang efektif ketika dihadapkan pada ribuan fitur teks yang mayoritas bernilai nol. Sementara itu, *SVM* mampu mendekati performa *Logistic Regression* dengan nilai akurasi 82,00% karena algoritma tersebut juga bekerja mencari batas linear (*hyperplane*) yang kuat pada dimensi tinggi.

Dengan demikian, hasil pengujian ini membuktikan bahwa penyelesaian kasus klasifikasi teks deskripsi lagu yang digabungkan dengan standardisasi fitur numerik tidak selalu membutuhkan model ensemble yang kompleks. Pendekatan model linear seperti *Logistic Regression* terbukti lebih efektif, stabil, dan efisien dalam memetakan popularitas lagu pada dataset Top 500 Songs.

IV. KESIMPULAN

Penelitian ini berhasil membandingkan performa empat algoritma klasifikasi, yaitu *Random Forest*, *Logistic Regression*, *Support Vector Machine (SVM)*, dan *Gradient Boosting* dalam mengklasifikasikan popularitas lagu menggunakan dataset *Top 500 Songs*. Proses klasifikasi dijalankan secara sistematis melalui tahapan *preprocessing* data, pembentukan label target biner (*Top 10* dan *Non Top 10*), ekstraksi fitur teks menggunakan pembobotan *Term Frequency-Inverse Document Frequency (TF-IDF)*, serta standardisasi fitur numerik menggunakan *StandardScaler*. Dataset kemudian dibagi menjadi data latih (*training data*) sebesar 80% dan data uji (*testing data*) sebesar 20%. Pengukuran kinerja dari setiap model diukur dan dievaluasi secara komprehensif menggunakan metrik *Accuracy*, *Precision*, *Recall*, *F1-Score*, dan visualisasi *Confusion Matrix*.

Berdasarkan hasil pengujian yang telah dilakukan, diperoleh data performa bahwa algoritma *Logistic Regression* menempati urutan tertinggi dengan nilai akurasi sebesar 84,00%. Posisi berikutnya diikuti oleh *Support Vector Machine (SVM)* dengan akurasi 82,00%, *Random Forest* sebesar 80,00%, dan *Gradient Boosting* yang menghasilkan akurasi terendah sebesar 78,00%. Hasil ini membuktikan bahwa

karakteristik data teks berdimensi tinggi dan renggang (*sparse matrix*) hasil transformasi TF-IDF lebih optimal jika dipetakan menggunakan pendekatan model linear seperti *Logistic Regression* dibandingkan model *ensemble* berbasis pohon keputusan.

Dengan tercapainya tujuan perbandingan performa ini, beberapa saran yang dapat diajukan untuk penelitian selanjutnya antara lain: memperluas cakupan dataset dengan jumlah sampel yang lebih besar, menerapkan pencarian parameter optimal (*hyperparameter tuning*) seperti *Grid Search* atau *Random Search* pada masing-masing algoritma, serta mengeksplorasi metode representasi teks yang lebih modern (seperti *Word2Vec*, *FastText*, atau *BERT*) demi meningkatkan kualitas dan akurasi klasifikasi ke tingkat yang lebih tinggi.

UCAPAN TERIMA KASIH

Puji syukur kehadirat Tuhan Yang Maha Esa atas rahmat dan karunia-Nya sehingga penelitian komparatif mengenai algoritma machine learning untuk klasifikasi popularitas lagu ini dapat dirampungkan dengan baik. Penulis menyampaikan rasa hormat dan terima kasih kepada institusi Universitas Duta Bangsa Surakarta yang telah memberikan dukungan serta lingkungan akademik yang kondusif selama penulis menempuh studi dan melaksanakan penelitian ini.

Penghargaan, apresiasi, dan rasa terima kasih yang mendalam secara khusus didedikasikan kepada seluruh anggota kelompok tim penulis, yaitu Fikko Nanda Ramadhani, Naaq Shal Ramadanu, Adit Ramadan, Muhammad Ridwan Martana Putra, dan Aprilisa Arum Sari. Penelitian ini dapat diselesaikan sepenuhnya berkat adanya kerja sama yang solid, komitmen yang luar biasa, serta sumbangsih pemikiran yang berharga dari setiap anggota kelompok mulai dari tahap pengumpulan data, *preprocessing*, eksperimen pemodelan secara mandiri, hingga penyusunan naskah ini.

Akhir kata, ucapan terima kasih yang tulus disampaikan kepada segenap panitia Seminar Nasional Teknologi Informasi dan Bisnis

(SENATIB) 2026 yang telah memberikan ruang, kesempatan, serta memfasilitasi publikasi ilmiah ini. Semoga hasil penelitian kelompok ini dapat memberikan kontribusi nyata dan menjadi referensi yang bermanfaat bagi perkembangan ilmu pengetahuan di bidang data mining dan klasifikasi teks.

REFERENSI

- [1] H. Allam, L. Makubvure, B. Gyamfi, K. N. Graham, and K. Akinwolere, "Text Classification : How Machine Learning Is Revolutionizing Text Categorization," 2025.
- [2] J. G. Muchori and M. P. Maina, "Review Article A Systematic Literature Review on Application of Text Embedding Techniques in NLP," vol. 13, no. December, pp. 45–61, 2025.
- [3] H. Barus, I. N. Fajri, and Y. Pristyanto, "Sentiment Classification Analysis of Tokopedia Reviews Using TF-IDF , SMOTE , and Traditional Machine Learning Models," vol. 9, no. 5, pp. 2552–2561, 2025.
- [4] Y. Zhao and Y. Che, "TF-IDF joint *SVM* model in library automation bias risk assessment system," 2026.
- [5] A. Fauzidan *et al.*, "ANALISIS KOMPARATIF EFEKTIVITAS PIPELINE DATA CLEANING BERBASIS," vol. XIV, no. 2, pp. 141–149, 2025.
- [6] Y. Yang, "An Evaluation of Statistical Approaches to Text Categorization/," vol. 90, pp. 69–90, 1999.
- [7] B. Practices, "Data Preprocessing and Feature Engineering for Data Mining :," 2025.
- [8] O. Hany, "500 Greatest Songs of All Time," 2020. [Online]. Available: <https://www.kaggle.com/datasets/omarhany/500-greatest-songs-of-all-time>
- [9] I. Rahmi, R. Efendi, N. A. Samat, F. Yanuar, and H. Yozza, "Essentials of Preprocessing Data in Improving *Logistic Regression* Performance Based on Rough Sets Theory : A Case Study of Stunting in West Sumatra , Indonesia," vol. 13, no. 1, pp. 124–139, 2025.
- [10] V. B. Lestari and C. A. Hutagalung, "Evaluation of TF-IDF Extraction Techniques in Sentiment Analysis of Indonesian-Language Marketplaces Using *SVM* , *Logistic Regression* , and Naive Bayes," no. 021, pp. 36–44.
- [11] F. Pedregosa, R. Weiss, and M. Brucher, "Scikit-learn : Machine Learning in Python," vol. 12, pp. 2825–2830, 2011.
- [12] B. Ramadhani and R. R. Suryono, "Komparasi Algoritma Naïve Bayes dan *Logistic Regression* Untuk Analisis Sentimen Metaverse," vol. 8, no. April, pp. 714–725, 2024, doi: 10.30865/mib.v8i2.7458