 ISSN:2302-8149
e-ISSN:2540-9719

Klasifikasi Bahasa Toraja, Batak dan Ambon menggunakan Metode Decision Tree dan Gradient Boost

Classification of Toraja, Batak and Ambon Languages using Decision Tree and Gradient Boost methods

[1]Bileam Mangalla*, [2]Suharyadi
[1,2]Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Satya Wacana
[1,2]Jl. Diponegori No.52-60, Salatiga, Kec. Sidorejo, Kota Salatiga, Jawa Tengah 50711
*e-mail: bileammangalla12@gmail.com

(***received*:** 24 February 2025, ***revised*:** 12 March 2025, ***accepted*:** 13 March 2025)

Abstrak

Dengan beragam suku, budaya, ras, dan agamanya, indonesia adalah satu negara dengan jumlah bahasa daerah terbanyak di dunia. Sering kali terjadi kesulitan dalam komunikasi karena banyaknya bahasa daerah di Indonesia. Ini terutama terjadi saat menyampaikan informasi atau percakapan dalam bentuk teks. Penelitian ini akan menunjukkan hasil identitas bahasa Toraja, Batak, dan Ambon menggunakan teknologi komputerisasi pembelajaran mesin. Metode yang digunakan termasuk *Decision Tree* dan *gradient Boost* untuk menghitung akurasi setiap metode. Hasil yang didapatkan dengan memakai metode *Decision tree* dan *Gradient Boost* adalah kedua metode efektif dalam melakukan identifikasi dengan nilai *accuracy* diatas 0,77 atau 77%. Namun dilihat dari *confusion matrix* metode *Gradient Boost* lebih efektif dengan nilai *accuracy 0,8106* atau 81,06%, dibandingkan dengan *Decision Tree* dengan nilai *accuracy* 0,7839 atau 78,39%. Hasil identifikasi dengan *Decision Tree* dan *Gradient Boost* keduanya efektif, dengan nilai akurasi lebih dari 0,77 atau 77%. Namun, berdasarkan *confusion matrix*, metode *Gradient Boost* lebih baik dengan nilai akurasi 81,06%, dibandingkan dengan *Decision Tree*, yang memiliki nilai akurasi 78,39%.

Kata kunci: klasifikasi, data mining, bahasa, decision tree, gradient boost

Abstract

With its rich diversity of ethnicities, cultures, races, and religions, Indonesia is one of the countries with the highest number of regional languages in the world. This linguistic diversity often leads to communication challenges, particularly when conveying information or engaging in textual conversations. This study aims to identify and classify the Toraja, Batak, and Ambon languages using machine learning-based computational methods. The techniques employed include Decision Tree and Gradient Boost algorithms to evaluate the accuracy of each model. The results demonstrate that both Decision Tree and Gradient Boost are effective in language identification, achieving accuracy rates above 77%. However, based on the confusion matrix analysis, the Gradient Boost method proved to be more effective, with an accuracy rate of 81.06%, compared to 78.39% achieved by the Decision Tree. These findings suggest that Gradient Boost offers better performance for classifying these regional languages.

Keywords: *classification, data mining, language , decision tree, gradient boost*

1 Pendahuluan

Indonesia adalah negara yang kaya akan keberagaman suku, budaya, ras, dan agama. Namun, perbedaan ini sering menjadi tentangan dalam komunikasi, karena setiap daerah memiliki bahasa dan dialek yang unik serta keberagaman[1] . Setiap daerah di Negara kesatuan Republik Indonesia (NKRI) terdiri dari daerah kabupaten maupun daerah kota. Setiap daerah di Indonesia memiliki perbedaan dalam berbagai aspek, seperti adat istiadat, suku, agama dan bahasa. Perbedaan ini dapat di temukan di berbagai daerah seperti Toraja yang memiliki bahasa sendiri yang di sebut bahasa Toraja, Batak dengan bahasa Bataknya, serta Ambon yang menggunakan bahasa Ambon.

 ISSN:2302-8149
e-ISSN:2540-9719

Bahasa merupakan alat komunikasi yang sangat penting bagi keberlangsungan hidup manusia. Sebagai element utama dalam kehidupan, bahasa memungkinkan manusia untuk berinteraksi dan menjalani kehidupan dengan teratur. Tanpa bahasa, keberlangsungan hidup manusia tidak akan berjalan dengan baik. Bahasa adalah sistem simbol bunyi yang dibuat oleh alat ucap dan bebas dan konvensional. Bahasa digunakan oleh suatu kelompok orang untuk berkomunikasi dan mengungkapkan pemikiran dan perasaan mereka[2]. Sebagai makhluk sosial, manusia sangat memerlukan komunikasi untuk berinteraksi, bertukar pendapat dan sebagai pemikiran. Percakapan atau perbincangan berperan sabagai alat berpikir, sehingga memahami bahasa daerah menjadi penting untuk mempermudah komunikasi. Namun, keberagaman bahasa membuat pemahaman manusia terhadap bahasa daerah menjadi terbatas, ditambah dengan kurangnya akses terhadap sumber belajar bahasa daerah. Karena itu, klasifikasi bahasa diperlukan untuk membuat manusia lebih mudah membedakan dan memahami bahasa setiap wilayah.

Klasifikasi adalah salah satu tugas penting dalam pengolahan data, dan merupakan salah satu teknik analisis data yang paling penting. Metode ini menggunakan data untuk menentukan nilainya dengan mengelompokkan fitur-fitur ke dalam kelas yang sesuai. Selama proses ini, vektor fitur yang tersedia telah diketahui kelasnya dan vektor fitur pelatihan digunakan untuk merancang proses pemilihan. Metode ini kenal sebagai pembelajaran yang diawasi dalam pengenalan pola[3]. Proses klasifikasi terdiri dari dua tahap, yaitu pelatihan data (data *training*) dan pengujian data (data *testing*). Dalam klasifikasi, ada banyak metode yang dapat digunakan, tetapi penelitian ini hanya menggunakan dua metode yaitu *Decision Tree* dan *Gradient Boost.*

Penelitian ini berjudul "Klasifikasi bahasa Daerah Toraja, Batak dan Ambon menggunakan Decision Tree dan Gradient Boosting". Dalam penelitian ini menggunakan tiga bahasa daerah, yaitu Toraja, Batak, dan Ambon. Ketiga bahasa ini memiliki karakteristik linguistik yang unik dan mewakili tiga wilayah yang berbeda di Indonesia, yaitu Sulawesi, Sumatra, dan Maluku. Bahasa Toraja memiliki sistem fonologi yang khas, bahasa Batak memiliki beberapa dialek serta sistem tulisan surat Batak, sedangkan bahasa Ambon merupakan bahasa kreol yang terbentuk dari pengaruh berbagai bahasa, termasuk melayu. Selain itu, banyaknya bahasa yang digunakan oleh orang Toraja, Batak, dan Ambon membuat sulit untuk orang asing atau masyarakat luar untuk mengenal asal usul bahasa yang digunakan. Dengan perkembangan teknologi saat ini, orang luar dapat membantu mengenal bahasa daerah yang digunakan. Penelitian ini bertujuan untuk mengembangkan model klasifikasi yang mampu membedakan bahasa daerah Toraja, Batak, dan Ambon berdasarkan fitur-fitur tertentu yang diidentifikasi melalui analisis data. Dengan menggunakan teknik klasifikasi, penelitian ini akan mengevaluasi efektivitas metode Decision *Tree* dan *Gradient Boosting* dalam mengelompokkan bahasa daerah. Selain itu, penelitian ini juga bertujuan untuk memvisualisasikan serta menganalisis perbedaan karakteristik bahasa menggunakan *Principal Component Analysis* (PCA), sehingga dapat memberikan pemahaman yang lebih dalam mengenai pola dari masing-masing bahasa.

Penelitian ini menggunakan sejumlah langkah untuk mengumpulkan data bahasa dari berbagai sumber, mengekstrak fitur yang relevan, dan menggunakan algoritma klasifikasi *Decision Tree* dan *Gradient Boosting*. Untuk memastikan bahwa model yang dikembangkan dapat mengenali dan mengklasifikasi bahasa daerah yang diuji dengan baik, data akan dibagi menjadi dua bagian yaitu data pelatihan dan data pengujian. Analisis akan dilakukan dengan membandingkan hasil klasifikasi dari kedua metode serta mengukur tingkat akurasi guna untuk menentukan metode yang paling optimal dalam membedakan bahasa daerah Toraja, Batak, dan Ambon.

Penelitian ini berkontribusi dalam pengembangan metode klasifikasi bahasa daerah yang dapat digunakan dalam bidang linguistik komputasional. Dengan adanya penelitian ini, diharapkan dapat memberikan wawasan baru dalam menggunakan algoritma machine learning untuk mengklasifikasi bahasa daerah. Hasil penelitian ini juga dapat digunakan sebagai referensi bagi penelitian selanjutnya yang berfokus pada pengolahan bahasa alami dan pengenalan bahasa daerah.

Penelitian ini memiliki beberapa manfaat yang signifikan, baik secara akademis maupun praktis. Secara akademis, penelitian ini berkontribusi dalam pengembangan ilmu linguistik komputasional, kusus dalam klasifikasi bahasa daerah menggunakan teknik machine learning. Hasil penelitian ini dapat menjadi pemrosesan bahasa alami. Secara praktis, penelitian ini dapat membantu masyarakat dan peneliti dalam memahami karakteristik bahasa daerah yang berbeda, sehingga dapat digunakan untuk mengembangkan aplikasi penerjemah, pengenalan bahasa, serta pelestarian bahasa

 ISSN:2302-8149
e-ISSN:2540-9719

daerah. Dengan adanya klasifikasi bahasa yang lebih sistematis, pemahaman dan pembelajaran bahasa daerah dapat menjadi lebih mudah dan terstruktur.

2 Tinjauan Literatur

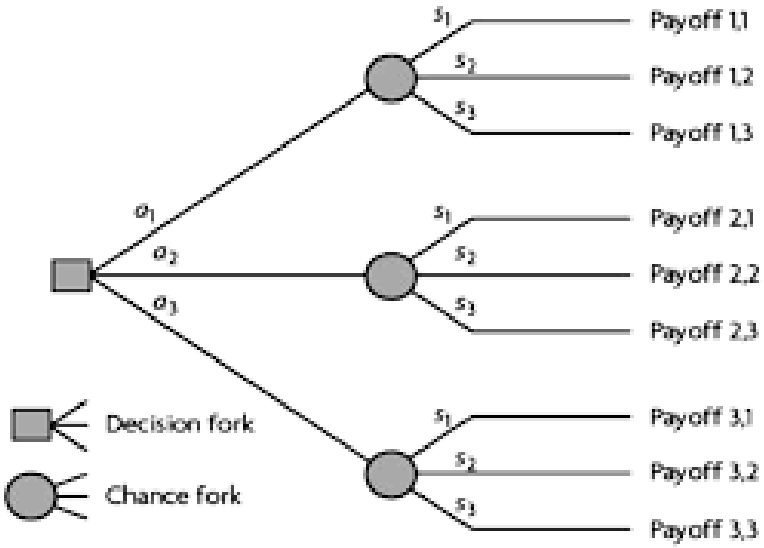
Pada penelitian ini menggunakan data mining. *Data mining* adalah proses mengelola data untuk menghasilkan informasi baru yang bermanfaat[4]. *Data mining* dapat dibagi menjadi beberapa jenis berdasarkan fungsinya: deskripsi estimasi, prediksi, klasifikasi, pengelompokan (*clustering*) dan asosiasi[5]. *Data mining* biasanya digunakan untuk membantu pengambilan keputusan di masa yang akan datang menggunakan informasi dari masa lalu[6]. Data mining dapat digunakan untuk mengklasifikasi dan memprediksi[7]. Klasifikasi adalah salah satu metode data mining[3]. Klasifikasi adalah proses pembuatan model atau fungsi untuk mengidentifikasi dan membedakan konsep atau kategori data dengan tujuan menentukan kelas objek yang labelnya belum diketahui[8]. Proses klasifikasi data meliputi dua tahap utama, yaitu Pembelajaran (*training*) dan pengujian (*testing*). Pada tahap pembelajaran, data latih dianalisis menggunakan algoritma klasifikasi, sedangkan pada tahap *testing*, data uji digunakan untuk mengevaluasi tingkat akurasi dari aturan klasifikasi yang di diterapkan[9]. Penelitian ini menggunakan *Decision Tree* dan *Gradient Boosting*, dua teknik klasifikasi yang telah banyak digunakan dalam penelitian sebelumnya.

Metode *convolution neural network (CNN)* dan *extreme gradient Boosting (XGBoost)* digunakan dalam studi ini untuk mengklasifikasikan penyakit pneumonia berdasarkan gambar radiografi. Mengukur tingkat akurasi dan membuat modal klasifikasi terbaik adalah tujuan utama penelitian ini. Dengan pembagian data 70:30, model *CNN* mencapai akurasi 88,73% dan *XGBoost* mencapai akurasi 91,12%. Hasil ini menunjukkan bahwa *XGBoost* lebih baik daripada *CNN* dalam mendeteksi penyakit pneumonia[10].

Dalam penelitian ini, pihak rektorat harus mempertimbangkan faktor-faktor berikut saat membuat keputusan untuk memberikan penghargaan kepada dosen yang berprestasi: perbandingan algoritme *Naive Bayes* dan *Decision Tree*(C4.5) dalam mengklasifikasi dosen yang berpresentasi. Penelitian ini akan di bandingkan hasil akurasi klasifikasi berdasarkan prestasi dosen. Fitur-fitur yandi pertimbangkan dalam klasifikasi meliputi publikasi internasional dan nasional, Tingkat Pendidikan, durasi studi doctoral, status sebagai rector kepala, serta lamanya sertifikasi dosen. Untuk mengkategorikan dosen berdasarkan prestasinya, digunakan algoritme *naïve bayes* dan *support vector machine*. Penelitian ini dilakukan di Universitas Sebelas Maret antara tahun 2018 hingga 2021. Pada penelitian ini, metode *Decision Tree* adalah yang terbaik karena akurasi pengujian *Cross-Validation* sepuluh kali lipat dengan metode Naive Bayes adalah 94,80%, sedangkan metode Decision Tree memiliki akurasi 95,80%[11].

Decision tree dianggap sebagai salah satu metode paling terkenal untuk *representasi* klasifikasi data pengklasifikasi[12]. Pohon Keputusan (*Decision tree*) merupakan salah satu metode yang cukup mudah diinterpretasikan oleh manusia[13]. *Decision Tree* adalah metode prediksi yang kuat dan terkenal. Pohon keputusan memiliki kemampuan untuk memeriksa data untuk mengidentifikasi hubungan tersembunyi antara variabel input dan potensi variabel target yang akan digunakan[7]. Metode keputusan pohon dalam pengelolaan data adalah metode klasifikasi yang menghasilkan klasifikasi berupa pohon keputusan berdasarkan sampel data yang dikumpulkan. Mengkonversi data menjadi struktur pohon yang terdiri dari keputusan dan aturan yang membantu proses pengambilan keputusan dikenal sebagai konsep pohon keputusan[13]. Pohon keputusan ini dapat berupa data *prediktor universal* atau *multivariat*[11]. Menghasilkan model prediksi yang spesifik dalam bentuk aturan yang jelas dan mudah digunakan adalah tujuan utama *decision tree*[14]. *Decision tree* berfungsi untuk mendeskripsikan setiap kelas guna menemukan pola atau fungsi, dengan tujuan melakukan klasifikasi atau memprediksi data yang belum memiliki kelas[9]. Simpul pohon terdiri dari tiga bagian: simpul akar (*roat* atau simpul), simpul perbangan/*internal* (cabang atau simpul *internal*), dan simpul daun (simpul daun)[9]. Gambar 1 menunjukkan bentuk umum pohon keputusan.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 1. Pola pohon keputusan
(Sumber: https://images.app.goo.gl/1xw9MkWuQwd57SE46)

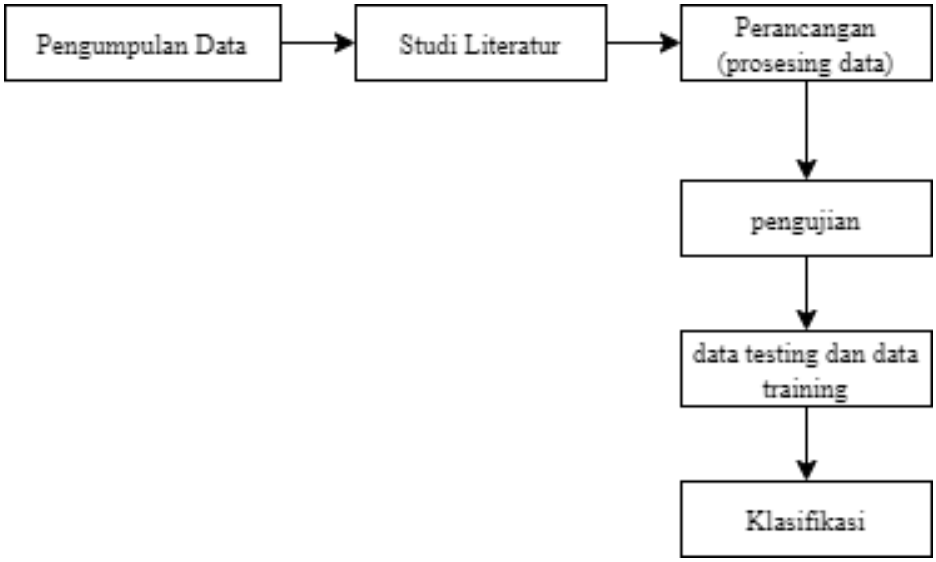
Decision tree adalah struktur berbentuk flowchart yang menyerupai pohon, seperti yang ditunjukkan pada Gambar 1 di atas, setiap simpul internal menunjukkan proses pengujian terhadap atribut dan setiap cabang menunjukkan hasilnya. Simpul daun juga dikenal sebagai simpul daun, berfungsi sebagai penanda kelas hasil klasifikasi. Simpul teratas dikenal sebagai root node, yang tidak memiliki beberapa cabang keluaran. Simpul internal juga dikenal sebagai internal node, menerima satu masukan dan menghasilkan beberapa keluaran, sedangkan simpul daun hanya menerima satu masukan dan tidak menghasilkan keluaran. Hasil akhir node cabang menunjukkan label kelas berdasarkan atribut yang telah diproses menjadi aturan(rule) [9].

Teknik busting, yang pertama kali diperkenalkan oleh Friedman pada tahun 2021, dapat dianggap sebagai model rata-rata awal yang digunakan untuk klasifikasi. Namun, itu juga dapat diterapkan pada *regresi*, seperti teknik *debugging* yang menggunakan *voting* untuk mencapai tujuan klasifikasi atau menghitung estimasi numerik rata-rata untuk setiap model keluaran; deskripsi model dengan tipe yang serupa, seperti pohon keputusan, adalah contoh tambahan[15]. *XGBoost* adalah teknik peningkatan kelompok berbasis pohon keputusan regresi dan menggunakan klasifikasi dan stres regresi satu pohon. Nilai salah atau salah dari pohon pertama (*feature*,Y) ke pohon berikutnya (*feature* |Y-Y1) dikurangi dengan menggabungkan beberapa pohon keputusan [11]. Algoritma GBM dapat mengoptimalkan proses sepuluh kali lebih cepat daripada algoritma GBM lain. Keakuratan hasil klasifikasi metode *Xboost* dipengaruhi oleh parameter yang dipilih. Penelitian ini menggunakan parameter *eta, Gamma, Max_depth, Min_child_weight, Subsample, dan Colsample_bytree*[15].

3 Metode Penelitian

Penulisan melewati berbagai langkah dalam proses identifikasi. Seperti pengumpulan data, penelitian literatur, perancangan, dan proses data. Gambar 2 menunjukkan langkah-langkah penelitian.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 2. Alur penelitian

Pengumpulan data adalah tahap pertama dalam proses identifikasi bahasa. Data ini dikumpulkan dari berbagai sumber, seperti media cetak dan internet, dan berupa teks bahasa sehari-hari dari bahasa Toraja, Batak, dan Ambon. Jumlah sampel yang tepat untuk mewakili populasi tidak diketahui, tetapi secara umum, lebih besar ukuran sampel, semakin besar kemungkinan hasilnya akan mewakili populasi dengan lebih akurat[16]. Untuk meningkatkatkan jumlah data yang memenuhi tujuan yang diinginkan, teks ini diubah menjadi teks yang lebih singkat. Sistem komputer akan menjadi lebih baik dalam menganalisis data seiring dengan jumlah data yang dimasukan. Data akan dimasukan secara manual ke *microsoft Excel*, dengan lebih dari 195.000 entri untuk setiap label. Agar metode *decision tree* dan *gradient boost* dapat diuji, data yang terdiri dari kalimat dalam bahasa Toraja, Batak, dan Ambon disebut sebagai data mentah.

Peneliti disarankan untuk menggunakan Metode pohon Keputusan dan *gradient boosting* saat melakukan klasifikasi. Tujuan dari tahap kedua penelitian ini adalah untuk mempermudah penelitian membandingkan hasil klasifikasi bahasa dengan kedua metode tersebut.

Pada tahap perancangan dilakukan pemrosesan data dengan menghilangkan tanda baca pada teks. Contohnya:

- *Kalimat: "LAMALEPA' MAMMA' DOLO"*
- *Pemrosesan : "Lamalepa Mamma Dolo" (menghilangkan tanda baca (')*

Data yang dientrikan disimpan dalam *microsoft Excel* dan dilabeli dengan kata "Tora" untuk wilayah Toraja, "batak" untuk wilayah Batak, dan "ambon" untuk bahasa Ambon. Agar sistem dapat menerjemahkan data saat diuji, data dibagi menjadi dua bagian: 80% data pelatihan dan 20% data uji. Setelah itu pada tahap identifikasi perhitungan uji akurasi, presisi, recal, dan skor F1 dilakukan. Sistem menunjukkan hasil dan performa dari Decision Tree dan Gradient Boosting, yang memungkinkan perbandingan.

4 Hasil dan Pembahasan

Bagian ini menyajikan hasil eksperimen dalam klasifikasi bahasa daerah Toraja, Batak, dan Ambon. Pembahasan mencakup analisis dataset, implementasi metode klasifikasi *Decision Tree* dan *Gradient Boosting*, serta evaluasi performa model berdasarkan matrik seperti akurasi, presisi, recall, dan skor F1. Selain itu, analisis terhadap fitur-fitur penting yang berpengaruh dalam proses klasifikasi turut dilakukan untuk memberikan wawasan lebih mendalam mengenai karakteristik unik dari masing-masing bahasa.

 ISSN:2302-8149
e-ISSN:2540-9719

Dataset

Data yang digunakan dalam pengembangan sistem ini terdiri dari bahasa Batak, Ambon, dan Toraja, yang diorganisasikan dalam bentuk file Ms, Excel. Data tersebut terbagi menjadi dua kolom utama: *cleantext*, yang berisi isi kalimat dalam setiap bahasa, dan *label*, yang berfungsi sebagai penanda identitas bahasa informasi yang dimasukan ke dalam file *Microsoft Excel* bisa di tinjau pada Tabel 1.

Tabel 1. Set data bahasa toraja, batak dan ambon

	CleanText	Label
0	iko	Tora
1	ambek	Tora
2	indok	Tora
3	sangmane	Tora
4	umbokoi batang kaleku	Tora
5	umpatorroanna padik	Tora
6	naporonno wai matangku	Tora
7	kutandai duka tu penammu	Tora
8	maparri untampena	Tora
9	maleku duka o rendengku	Tora
10	ale pi mana	Ambon
11	dong su datang	Ambon
12	jang lai beta bale	Ambon
13	ini oto bagini bae	Ambon
14	ale pung rumah jauhkah	Ambon
15	sonde usah takut semua bae bae	Ambon
16	beta su capek kerja seharian	Ambon
17	ale mau pi pantai ka	Ambon
18	dong dua basodara	Ambon
19	beta su lama seng pulang kampung	Ambon
20	ise goarmu	Batak
21	unang lao ho	Batak
22	unang marmeam ho disi	Batak
23	naeng modom au	Batak
24	anak siakkangan	Batak
25	burjuhon damang namar sikkolai	Batak
26	Burjuhon ma karejo	Batak
27	Dang adong dope burjuhon	Batak
28	ateatehon tadinghononmu	Batak
29	Dang hea loja au	Batak

Microsoft Excel dibagi menjadi *CleanText* untuk menerjemahkan semua bahasa daerah dan label untuk menerjemahkan semua bahasa daerah selama proses entri data manual. Jumlah data untuk bahasa Toraja, Batak, dan Ambon lebih dari 3.000. Data yang dimasukkan tanpa tanda baca[17]. Jumlah data minimal yang harus dimasukkan kedalam sistem adalah tiga ribu data. Tujuannya adalah agar sistem dapat mempelajari pola, memproses kemungkinan, dan membagi data sesuai dengan rumus penelitian (membagi 20% data uji dan 80% data pelatihan).

Decision Tree Classifier

Kode 1 untuk *Decision Tree Classifier* digunakan untuk mengevaluasi tingkat akurasi sistem dengan menggunakan Confusion Matrix, seperti yang ditunjukkan pada gambar.

Kode 1. Program decesion tree

```
labels = ['batak', 'ambon', 'tora']
cm_Model_dt = confusion_matrix(y_test, y_pred)
```

 ISSN:2302-8149
e-ISSN:2540-9719

```
fig_dt = plt.figure(figsize=(9,9))
x_dt = fig_dt.add_subplot(111)
sns.heatmap(cm_Model_dt, annot=True, fmt=".3f", linewidths=.5, square = True,
cmap = 'Blues_r')
plt.ylabel('Nilai Asli')
plt.xlabel('Label Prediksi')
x_dt.set_xticklabels(labels)
x_dt.set_yticklabels(labels)
title = 'Decision Tree Model Accuracy Score = '+
str(round(accuracy_score_dt*100,2)) +"%"
plt.title(title, size = 15)
```

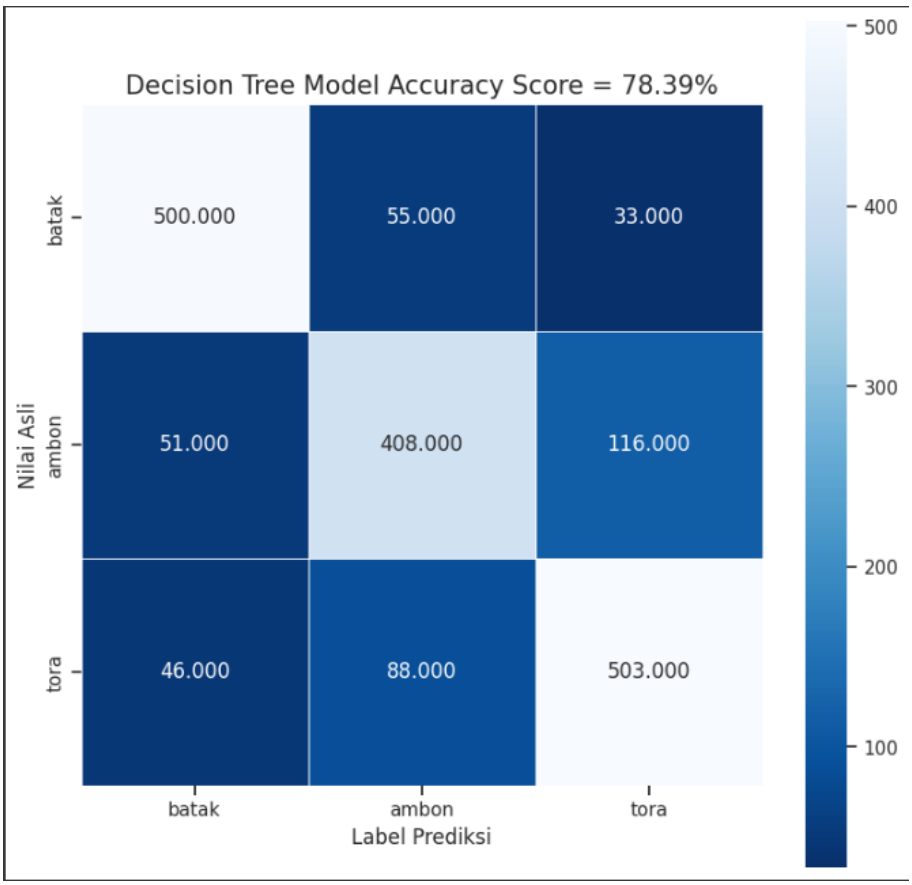
Dengan menghitung jumlah prediksi yang tepat, confusion matrix menyajikan ringkasan hasil prediksi dalam klasifikasi. Dianalisis kemudian berdasarkan setiap kelas atau label bahasa. Skor akurasi dikalikan dengan 100,2 untuk mendapatkan nilai akurasi.

Decision Tree Classifier Results

Tabel 2 menunjukan hasil perhitungan confusion matrix, dan Gambar 3 menunjukkan *heatmap comfusion Matrix Decision Tree.*

Tabel 2. Confusion matrix decision tree

	Precision	Recall	F1-Score	Support
Ambon	0.84	0.85	0.84	588
Batak	0.74	0.71	0.72	575
Tora	0.77	0.79	0.78	637
accuracy			0.78	1800
macro average	0.78	0.78	0.78	1800
weighted average	0.78	0.78	0.78	1800



 ISSN:2302-8149
e-ISSN:2540-9719

Gambar 3. Heatmap comfusion matrix decision tree

Hasil analisis korelasi secara keseluruhan, yang didasarkan pada uji matrik korelasi yang menggunakan 20% data uji, disajikan dalam Tabel 2. Analisis ini menilai presisi ingatan, dan skor F1 dengan akurasi 0,7839 atau 78,39%.

Identifikasi Decision Tree

Setiap kalimat yang ditulis dalam berbagai bahasa daerah akan diidentifikasi dengan menggunakan sistem yang didasarkan pada bahasa daerah yang paling sering digunakan. Kemampuan sistem untuk membedakan bahasa Batak, Ambon, Dan Toraja ditunjukkan pada Gambar 4. Misalnya kalimat input seperti "Tuat tu taru Martumba", "beta" dan "dipasiala" berhasil diidentifikasi oleh sistem sebagai bahasa Batak, Ambon, dan Toraja.

```
[ ]  predictions = pipeline_dct.predict(["Tuat tu toru martumba","beta","dipasiala"])
     predictions

     array(['batak', 'ambon', 'tora'], dtype=object)
```

Gambar 4. Prediksi bahasa decision tree

Gradient Boost

Kode 2 menampilkan *Gradient Boost Classifier*, yang digunakan untuk mengevaluasi Confusion Matrix.

Kode 2. Program gradient boost

```
labels = ['batak', 'ambon', 'tora']
cm_Mdl_gb = confusion_matrix(y_test,y_pred)
fig = plt.figure(figsize=(9,9))
x_gb = fig.add_subplot(111)
sns.heatmap(cm_Mdl_gb, annot=True, fmt=".3f", linewidths=.5, square = True, cmap
= 'Blues_r')
plt.ylabel('True Label')
plt.xlabel('Label Prediksi')
x_gb.set_xticklabels(labels)
x_gb.set_yticklabels(labels)
title = 'Gradient Boost Model Performance Score = '+
str(round(accuracy_score_gb*100,2)) +"%"
plt.title(title, size = 15)
```

Dengan menggunakan pembagian data uji sebesar 0,22 atau sekitar 20% dari data, dan perhitungan skor ketepatan *gradient boost* dikalikan dengan 100,2.

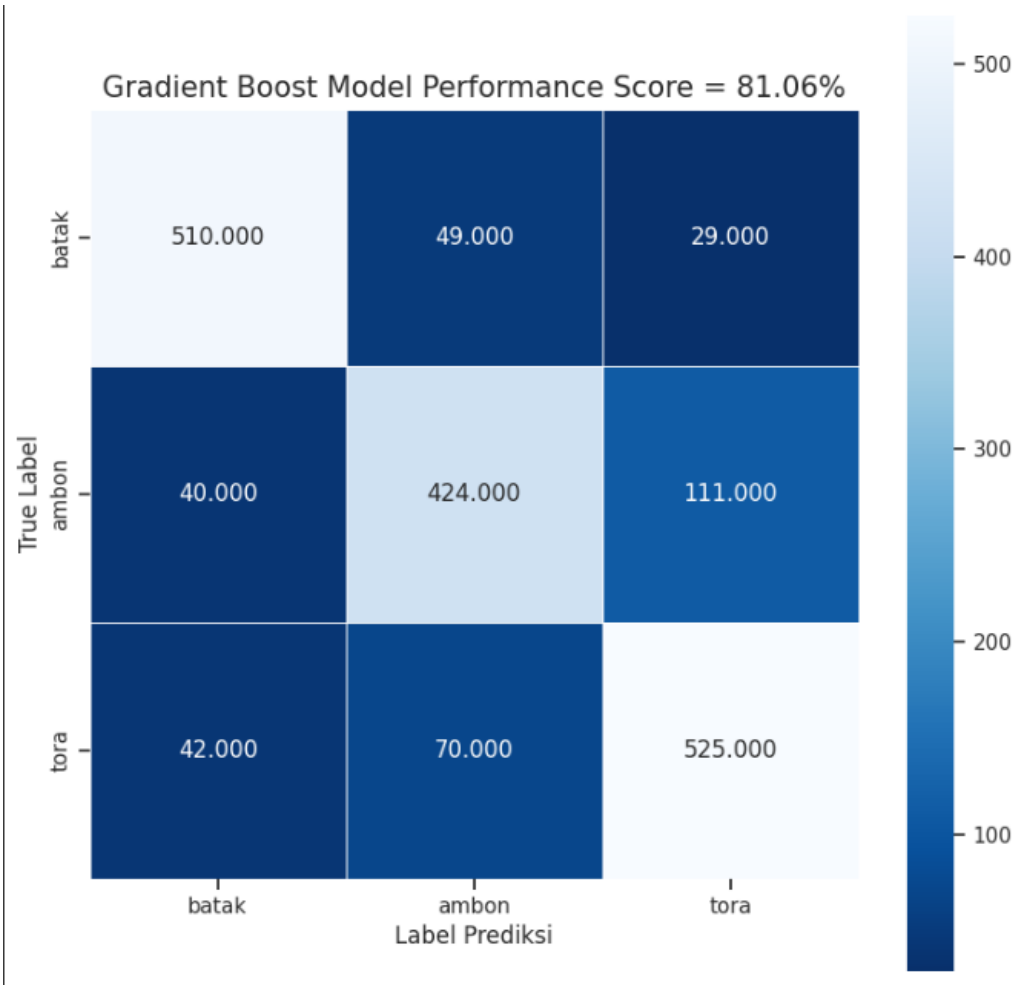
Gradient Boost Result

Tabel 3 menunjukan hasil perhitungan confusion matrix, dan Gambar 5 menunjukkan *heatmap gradient Boost Classifier*.

Tabel 3. Output confusion matrix gradient boost

	Precision	Recall	F1-Score	Support
Ambon	0.89	0.87	0.86	588
Batak	0.78	0.75	0.76	575
Tora	0.79	0.82	0.81	637
accuracy			0.81	1800
macro avg	0.81	0.81	0.81	1800
weighted avg	0.81	0.81	0.81	1800

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 5. Heatmap comfusion matrix gradient boost

Berdasarkan data set keseluruhan perhitungan keakuratan dapat diketahui nilai *accuracy* dari *gradient boost* yaitu sebesar 81.06%.

Identifikasi Gradient Boost

Sistem memasukan kalimat untuk mengetahui apakah mereka diterjemahkan sesuai dengan bahasa yang telah ditentukan. Sebagai contoh, kalimat "toguan urat ni padang","dong su", dan "sibayo", masing-masing diartikan sebagai, "batak","ambon", dan "tora", seperti yang ditunjukan pada Gambar 6.

```
predictions = pipeline_gb.predict(["toguan urat ni padang","dong su","sibayo"])
predictions

array(['batak', 'ambon', 'tora'], dtype=object)
```

Gambar 6. Prediksi bahasa gradient boost

Perbandingan Tingkat Akurasi

Tabel 4 menunjukkan hasil dari metode *Decision Tree* dan *Gradient Boosting,* yang didasarkan pada hasil analisis perbandingan nilai pada Confusion Matrix.

Tabel 4. Perbandingan tingkat akurasi

Metode	Tingkat Keakuratan
Decision Tree	78.39%
Gradient Boost	81.06%

 ISSN:2302-8149
e-ISSN:2540-9719

5 Kesimpulan

Berdasarkan Hasil penelitian yang membandingkan metode Pohon keputusan dengan *Gradient boosting*, metode *Gradient Boosting* menunjukkan kinerja menerjemahkan bahasa yang lebih baik dengan tingkat akurasi 80,06%. Sebaliknya *decision tree* tidak mampu mencapai tingkat akurasi tersebut dan hanya mencatat akurasi di bawah 80%. Analisis menggunakan confusion matrix, yang meliputi *precision, recal,* dan *f1-score,* menunjukan bahwa akurasi metode *gradient boost* lebih tinggi dibandingkan *decision tree*. *Gradient boost* mencapai akurasi sebesar 0,8106 atau 81,06%, sementara *decision tree* hanya mencapai tingkat akurasi 0,7839 atau 78,39%. Oleh karena itu, dapat disimpulkan bahwa *Gradient Boosting* melakukan identifikasi dengan lebih baik daripada *Decision Tree*.

Referensi

[1] D. Tuhenay and E. Mailoa, "Perbandingan Klasifikasi Bahasa menggunakan Metode Naïve Bayes Classifier (NBC) dan Support Vector Machine (SVM) Comparison of Language Classification using Naive Bayes Classifier (NBC) and Support Vector Machine (SVM) Method," *JIKO (Jurnal Inform. dan Komputer)*, vol. 4, no. 2, pp. 105–111, 2021, doi: 10.33387/jiko.

[2] I. N. T. Astawa, "Bahasa Indonesia sebagai Alat Pemersatu Bangsa," *Dharma Sastra J. Penelit. Bhs. dan Sastra Drh.*, vol. 2, no. 1, pp. 72–82, 2022, doi: 10.25078/ds.v2i1.940.

[3] A. P. Wibawa, M. Guntur, A. Purnama, M. Fathony Akbar, and F. A. Dwiyanto, "Metode-Metode Klasifikasi," *Pros. Semin. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 1, pp. 134–138, 2018.

[4] D. Ariyanto, "Data Mining menggunakan Algoritma K-Means untuk Klasifikasi Penyakit Infeksi Saluran Pernafasan Akut," *J. Sistim Inf. dan Teknol.*, vol. 4, pp. 13–18, 2022, doi: 10.37034/jsisfotek.v4i1.117.

[5] A. M. Argina, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 29–33, 2020, doi: 10.33096/ijodas.v1i2.11.

[6] G. M. Momole, "Perbandingan Naïve Bayes dan Random Forest dalam Klasifikasi Bahasa Daerah," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 2, pp. 855–863, 2022, doi: 10.35957/jatisi.v9i2.1857.

[7] T. W. Pratiwi and T. Arifin, "Optimasi Decision Tree menggunakan Particle Swarm Optimization untuk Klasifikasi Kesuburan pada Pria," *Sistemasi*, vol. 10, no. 1, p. 13, 2021, doi: 10.32520/stmsi.v10i1.967.

[8] Anggi Trifani, Agus Perdana Windarto, and Hendry Qurniawan, "Penerapan Data Mining Klasifikasi C4.5 dalam menentukan Tingkat Stres Mahasiswa Akhir," *Jural Ris. Rumpun Ilmu Tek.*, vol. 1, no. 2, pp. 91–105, 2022, doi: 10.55606/jurritek.v1i2.414.

[9] A. Solihin, D. I. Mulyana, and M. B. Yel, "Klasifikasi Jenis Alat Musik Tradisional Papua menggunakan Metode Transfer Learning dan Data Augmentasi," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 5, no. 2, pp. 36–44, 2022, doi: 10.47970/siskom-kb.v5i2.279.

[10] A. E. Putra, K. Kartini, and A. P. Sari, "Metode Convolutional Neural Network dan Extreme Gradient Boost untuk Mengklasifikasi Penyakit Pneumonia," *JASIEK (Jurnal Apl. Sains, Informasi, Elektron. dan Komputer)*, vol. 6, no. 1, pp. 33–40, 2024, doi: 10.26905/jasiek.v6i1.11464.

[11] A. Supriyadi, "Perbandingan Algoritma Naive Bayes dan Decision Tree(C4.5) dalam Klasifikasi Dosen Berprestasi," *Gener. J.*, vol. 7, no. 1, pp. 39–49, 2023, doi: 10.29407/gj.v7i1.19797.

[12] B. Charbuty and A. Abdulazeez, "Classification based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, 2021, doi: 10.38094/jastt20165.

[13] A. H. Nasrullah, "Implementasi Algoritma Decision Tree untuk Klasifikasi Produk Laris," *J. Ilm. Ilmu Komput.*, vol. 7, no. 2, pp. 45–51, 2021, doi: 10.35329/jiik.v7i2.203.

[14] R. Nursyahfitri, A. N. Maharadja, R. A. Farissa, and Y. Umaidah, "Klasifikasi Penentuan Jenis Obat menggunakan Algoritma Decision Tree," *J. Inform. Polinema*, vol. 7, no. 3, pp. 53–60,

 ISSN:2302-8149
e-ISSN:2540-9719

2021, doi: 10.33795/jip.v7i3.629.
[15] R. Siringoringo, R. Perangin Angin, and B. Rumahorbo, "Model Klasifikasi Genetic-Xgboost dengan T-Distributed Stochastic Neighbor Embedding pada Peramalan Pasar," *J. TIMES*, vol. 11, no. 1, pp. 30–36, 2022, doi: 10.51351/jtm.11.1.2022672.
[16] T. Dengan, C. Algoritma, D. M. Musa, D. Sakti, K. A. Shantiony, and S. K. Putri, "Penerapan Data Mining untuk Klasifikasi Data Penjualan Pakan Ternak Penerapan Data Mining untuk Klasifikasi Data Penjualan Pakan Ternak Terlaris dengan Algoritma C4 . 5," no. March, 2024, doi: 10.37012/jtik.v10i1.1985.
[17] Y. Azhar, A. K. Firdausy, and P. J. Amelia, "Perbandingan Algoritma Klasifikasi Data Mining untuk Prediksi Penyakit Stroke," vol. 5, no. 2, pp. 191–197, 2022.