Jurnal Teknik Informatika (JUTIF)
Vol. 4, No. 5, October 2023, pp. 1249-1255

DOI: https://doi.org/10.52436/1.jutif.2023.4.5.1459
p-ISSN: 2723-3863
e-ISSN: 2723-3871

CLASSIFICATION OF REGIONAL LANGUAGES USING METHODS GRADIENT BOOTS AND RANDOM FOREST

Eva Sapan Patasik[*1], Sri Yulianto[2]

[1,2]Informatika, Fakultas Teknologi Informasi, Universitas Kristen Satya Wacana, Indonesia
Email: [1]672018111@student.uksw.edu, [2]sri.yulianto@uksw.edu

(Article received: October 09, 2023; Revision: October 31, 2023; published: November 13, 2023)

Abstract

Indonesia is one of the countries that has the most regional languages in the world, ranking second most. The large number of regional languages that are owned makes it difficult for people between regions to recognize the origins of the regional language, so the author aims to conduct research by identifying a regional language. Identifying a language using data mining, one of the data mining techniques is classification. Classification is a technique used to find the value of data. Classification will build a model from samples of data into groups of the same type. There are two classification methods used in this research, namely gradient boots and random forest, where the two methods will be compared using regional language data from Java, Nias and Toraja. The results of calculating the accuracy values for the two methods used are quite good in classifying languages with results of an accuracy level of 0.8 or 80%, where the results of the gradient boots research have an accuracy value of 0.8850 or 88.5%, while the random forest method has an accuracy value. random forest is lower, namely 0.8794 or 87.94%, so in this study the gradient boots method is the best method.

Keywords: *Classification, Gradient Boots, Language, Random Forest.*

KLASIFIKASI BAHASA DAERAH MENGGUNAKAN METODE GRADIENT BOOTS DAN RANDOM FOREST

Abstrak

Indonesia merupakan salah satu negara yang memiliki bahasa daerah terbanyak di dunia yang mempunyai urutan posisi kedua terbanyak. Banyaknya bahasa daerah yang dimiliki membuat masyarakat antar daerah memiliki kesulitan untuk mengenali asal-usul bahasa daerah tersebut, sehingga penulis mempunyai tujuan melakukan penelitian dengan mengidentifikasi suatu bahasa daerah. Mengidentifikasi suatu bahasa menggunakan data mining, salah satu teknik data mining adalah klasifikasi. Klasifikasi merupakan teknik yang digunakan untuk menemukan nilai dari suatu data, klasifikasi akan membangun suatu model dari sambel data menjadi suatu kelompok yang sejenis. Terdapat dua metode klasifikasi yang digunakan dalam penelitian ini yaitu *gradient boots* dan *random forest*, dimana kedua metode akan dilakukan perbandingan menggunakan data bahasa daerah Jawa, Nias, dan Toraja. Pada hasil perhitungan nilai *accuracy* pada kedua metode yang digunakan cukup bagus dalam melakukan klasifikasi bahasa dengan hasil tingkat *accuracy* 0,8 atau 80% dimana hasil penelitian *gradient boots* memiliki nilai *accuracy* sebesar 0,8850 atau 88,5% sedangkan metode *random forest* memiliki nilai *accuracy random forest* lebih rendah yaitu 0,8794 atau 87,94%, sehingga dalam penelitian ini metode *gradient boots* merupakan metode terbaik.

Kata kunci: *Bahasa, Gradient Boots, Klasifikasi, Random Forest.*

1. PENDAHULUAN

1.1. Latar Belakang

Indonesia merupakan negara kepulauan yang memiliki suku bangsa, budaya, dan Bahasa. Kita memiliki Bahasa pemersatu bangsa yaitu Bahasa Indonesia. Setiap daerah memiliki Bahasa yang berbeda-beda yang digunakan sebagai media komunikasi, Penggunaan Bahasa daerah telah mengalami penurunan penggunaan dalam Bahasa komunikasi sehari-hari[1]. Jawa, Nias dan Toraja memiliki bahasa daerahnya masing-masing, Bahasa daerah Jawa merupakan Bahasa dengan penutur terbanyak di Indonesia yang digunakan oleh suku Jawa yang meliputi Jawa Tengah, Yogyakarta dan Jawa Timur. Bahasa daerah Toraja digunakan oleh suku toraja yang tersebar di kabupaten Toraja Utara dan Kabupaten Mamasa. Bahasa Toraja masih

memiliki beberapa dialek di daerah kabupaten Tana Toraja, yang dibagi atas tiga dialek yaitu Makale-Rantepao, Salaputti-Bonggakaradeng dan Sillanan-Gandangbatu. Sedangkan Bahasa Nias, atau Li Niha dalam bahasa aslinya, adalah bahasa yang dipergunakan oleh penduduk kepulauan Nias . Bahasa Nias merupakan salah satu bahasa di dunia yang masih belum diketahui persis dari mana asalnya.

Bahasa daerah merupakan salah satu kekayaan yang dimiliki oleh Bangsa Indonesia. Berdasarkan hasil penelitian Badan Pengembangan, Pembinaan Bahasa, Kementerian Pendidikan dan Kebudayaan bahasa daerah yang sudah teridentifikasi dan tervalidasi jumlahnya ada 652 bahasa dari 2.452 daerah[2]. Banyak bahasa daerah yang dimiliki, membuat masyarakat antar daerah sulit untuk mengenal asal usul daerah yang dipakai. kemampuan mengetahui bahasa setiap daerah sangat kurang, selain itu kurangnya sarana yang dapat diakses untuk mengetahui bahasa setiap daerah. Oleh karena itu penulis melakukan suatu klasifikasi bahasa, dimana klasifikasi ini memudahkan kita untuk mengetahui dan membedakan suatu bahasa daerah, dengan menggunakan metode *Gradient Boots* dan *Random Forest* sebagai teknik klasifikasi untuk melakukan prediksi.

Salah satu teknik-teknik Data Mining adalah Klasifikasi. Klasifikasi dalam data mining merupakan metode pembelajaran untuk memprediksi nilai dari sekelompok atribut dan teknik klasifikasi digunakan untuk menemukan nilai dari suatu data[3]. Klasifikasi akan membangun suatu model dari sampel data menjadi satu kelompok yang sejenis. Terdapat 2 metode klasifikasi yang digunakan dalam penelitian ini, yaitu *Gradient Boots* dan *Random Forest*. *Gradient Boots* merupakan algoritma *machine learning* yang menggunakan *ensamble* dari *decision tree* untuk memprediksi nilai[4]. Sedangkan metode *Random Forest* adalah metode algoritma *dalam machine learning* yang digunakan untuk pengklasifikasian data set dalam jumlah besar, algoritma dan merupakan salah satu teknik klasifikasi untuk melakukan prediksi[5].

Berdasarkan permasalahan diatas, adapun tujuan penulis dalam penelitian ini yaitu Klasifikasi Bahasa Daerah Menggunakan Metode *Gradient Boots* dan *Random Forest* untuk mengetahui algoritma mana yang paling efektif untuk memprediksi keakuratan setiap metode.

1.2. Tinjauan Pustaka

Pada Penelitian kali ini penulis akan menyertakan beberapa karya ilmiah yang berkaitan dengan hasil karya ilmiah yang telah dibuat oleh penulis. Penulis menggunakan teknik yaitu data mining, dimana terdiri dari metode *Gradient Boots* Dan *Random Forest*. Dimana metode *Gradient Boots* dan *Random Forest* itu sangat sering dipergunakan dalam hal klasifikasi. Adapun beberapa penelitian yang membahas atau yang berkaitan pada metode tersebut sebagai berikut:

Berdasarkan hasil penelitian dan perbandingan yang dilakukan pada studi kasus klasifikasi bahasa Toraja, Halmahera, Kalimantan metode *gradient boots* cukup bagus digunakan dalam klasifikasi bahasa, karena memiliki tingkat akurasi diatas 0.6 atau 60%. Dilihat dari nilai *Precision*, *Recall*, dan *Fl-Score* menyatakan bahwa algoritma *gradient boots* merupakan algoritma yang memiliki tingkat keakuaratan yang tinggi dengan tingkat akurasi sebesar 0,6525 atau 62,25%[6]. Prediksi kemungkinan diabetes pada tahap awal menggunakan metode algoritma klasifikasi *random forest*. Penelitian ini bertujuan untuk Percobaan yang dilakukan pada dataset diabetes hospital, hasil menentukan bahwa algoritma random forest telah bekerja dengan akurasi terbaik dengan akurasi yang dicapai sebesar 97,88%[7]. Perbandingan metode *random forest* untuk klasifikasi tingkat penyakit hepatitis C pada *imbalance class data*, hasil perbandingan dapat dilihat pada nilai akurasi yang cukup baik yaitu 74,79%, hasil tersebut akan berpegaruh pada variabel hasil uji laboratorium seperti sel darah, jumlah enzim ALT, serta jumlah HCVRNA dalam tubuh pasien[8].

Berdasarkan hasil analisis klasifikasi bahasa daerah Halmahera, Kalimantan, Toraja dengan menggunakan metode *Gradient Boots* sistem dapat melakukan klasifikasi dengan baik, dimana dengan menggunakan metode *Gradient Boots* memperoleh hasil presentasi akurasi sebesar 0,6518 atau 65.18%[9]. Prediksi curah dari hari-hari berikutnya, dengan menggunakan metode klasfikasi *random forest*, sistem prediksi cuaca yang telah kami buat mendapatkan tingkat akurasi tertinggi diperoleh klasifikasi *random forest* dengan resampling yakni sebesar 95.59%[10].

Dari perhitungan menggunakan data penduduk desa yang diperoleh dari kantor kelurahan, metode *gradient boots* untuk klasifikasi penerima program bantuan sosial menghasilkan nilai yang tinggi dibandingkan dengan metode yang lain yaitu 93.15%[11]. Penelitian ini telah menguji algoritma *random forest* dengan optimasi GA dan Bagging, belum mampu meningkatkan akurasi algoritma *random forest* untuk klasifikasi set data Bank. Menggunakan optimasi ataupun tidak akurasi yang diperoleh adalah 88,30%, belum memberikan nilai akurasi yang cukup baik untuk klasifikasi data Bank[12]. Penerapan metode *random forest* untuk analisis resiko pada dataset *peer to peer* lending, penggunaan keseluruhan data tanpa adanya penyesuaian memiliki nilai akurasi tertinggi yaitu dengan nilai akurasi 0.924012 dapat dikatakan bahwa penggunaan metode *random forest* terhadap dataset pinjaman sangat cocok implementasinya dalam menghasilkan model baru[13].

Data mining adalah proses yang menggunakan teknik statistic, matematika, kecerdasan buatan,

machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar. Data mining disebut *knowledge discovery* adalah proses pengambilan pola pada data yang akan diproses lalu output tersebut berupa informasi yang penting[14]. Dimana pada penelitian kali ini menggunakan sebuah teknik data mining yang dapat mengimplementasikan metode *Gradient Boots* dan *Random Forest*, yang berguna untuk membandingkan tingkat akurasi yang paling maksimal dari setiap masing-masing metode.

Gradient Boots merupakan algoritma yang menggunakan teknik *ensamble* dari *decision tree*, algoritma dapat menyelesaikan persoalan klasifikasi dan prediksi data. Dimana teknik boosting ini dapat dilihat sebagai metode model rata-rata yang awalnya 24 dirancang untuk metode klasifikasi tetapi juga dapat diaplikasikan pada metode regresi, seperti dengan metode bagging yang memanfaatkan voting untuk tujuan mengklasifikasikan atau menghitung rata-rata estimasi numerik untuk output model individu tunggal[15]. Persamaan lainnya yaitu menggabungkan model yang mempunyai jenis yang sama, seperti pada metode decision tree. XGBoost adalah algoritma *machine learning* system *pada tree boosting* yang dapat berkembang atau dibuat untuk sistem tree yang lebih besar XGBoost juga merupakan suatu lanjutan dari *Gradient Boosting* merupakan *ensemble* dari model yang digunakan pada *decision tree* untuk dikembangkan dan mendapatkan running yang lebih cepat walaupun dalam memproses yang lebih besar[16].

Random Forest adalah algoritma yang belajar pada mesin yang memiliki banyak pohon keputusan, dengan kombinasi dari model bagging dan *random sub spaces*, metode *Random Forest* ini telah membuktikan tingkat keberhasilan dalam melakukan prediksi dan klasifikasi pada beberapa tahun terakhir dan menjadi salah satu algoritma *machine learning* terbaik yang dapat digunakan pada berbagai bidang. Selain itu algoritma *Random Forest* memiliki beberapa kelebihan yaitu dapat meningkatkan hasil akurasi jika terdapat data yang hilang, dan untuk resisting outliers, dapat menghasilkan eror yang relatif rendah, selain itu *Random Forest* mempunyai proses seleksi fitur dimana mampu mengambil fitur terbaik sehingga dapat meningkatkan performa terhadap model klasifikasi. adanya fitur tersebut *Random Forest* dapat bekerja pada big data dengan parameter yang kompleks secara efektif[17]..

2. METODE PENELITIAN

Dalam proses identifikasi bahasa penulis menggunakan beberapa tahapan yang dapat dilakukan oleh penulis, yaitu sebagai berikut:

Gambar 1. Tahapan Penelitian

Tahapan penelitian pada Gambar 1, dapat dijelaskan sebagai berikut:

1. Tahap awal dimulai dengan identifikasi masalah yang akan diangkat. Dimana tahap ini merupakan fase awal untuk mengetahui masalah yang terjadi pada data bahasa dan mencari solusi yang tepat untuk permasalahan yang ada.
2. Setelah melakukan identifikasi masalah tahap yang dilakukan selanjutnya yaitu Studi literatur dimana studi literatur merupakan suatu penyelesain untuk memecahkan persoalan dengan melakukan atau menelusuri berbagai informasi atau sumber-sumber tulisan atau jurnal yang pernah dibuat sebelumnya, buku, beserta referensi dan sumber yang terpercaya.
3. Tahap ketiga yaitu diskusi dan konsultasi dimana dalam pembuatan suatu penelitian dan penyusunan pada tugas akhir. Tahap ini membutuhkan diskusi atau konsultasi yang dilakukan dengan dosen pembimbing atau bisa juga dengan pihak lain yang ahli dalam bidang tersebut.
4. Tahap keempat pengumpulan data, dimana dilakukan dengan mengumpulkan data dari sumber internet, dan berdasarkan bahasa sehari atau percakapan yang dilakukan yaitu bahasa yang diambil dari bahasa daerah Jawa, Nias, dan Toraja. Data ini dibuat sebuah teks atau kalimat yang lebih singkat dan data ini dikumpulkan di dalam ms. excel dengan total jumlah data keseluruhan 3 bahasa yaitu ada 9000 dimana setiap bahasa daerah atau label ini mempunyai masing-masing minimal 3000.
5. Tahap akhir melakukan pengolahan data dimana penulis akan mengumpulkan keseluruhan data yang telah didapatkan kemudian melakukan tahap pengolahan seperti dengan cara penginputan ke dalam Ms. Excel dengan manual. Kemudian melakukan cleansing data dimana cleansing data sendiri ini adalah proses untuk memeriksa atau mengecek keseluruhan data dalam database dan menghapus atau memperbarui suatu informasi yang tidak akurat,

atau tidak lengkap. Melakukan cleansing data ini supaya data yang kita punya memiliki data yang kualitas yang terbaik. Setelah data sudah siap di pakai.

Dataset merupakan data yang digunakan penelitian ini adalah data bahasa daerah dari bahasa Jawa, Nias dan Toraja. Bahasa dikumpul dari berbagai bahasa sehari – hari, kemudian data tersebut digabungkan dan dimasukkan kedalam file Ms. Excel yang dipisahkan menjadi cleantext dan label untuk mengisi data kalimat dari masing – masing bahasa dan mengisi identitas daerah atau nama daerah bahasa tersebut. Nama daerah yang akan dimasukkan ke dalam label yaitu jawa, nias dan tora sebagai bahasa daerah Toraja. Data akan di cleansing terlebih dahulu dari tanda baca maupun angka, proses penginputan data dilakukan secara manual dari sumber data dengan menginput kalimat bahasa daerah pada cleantext dan nama daerah pada label. Jumlah keseluruhan data yang dimiliki adalah 9000 data dimana setiap bahasa terdiri dari 3000 data.

3. HASIL DAN PEMBAHASAN

Gradient boots classifier tahap codingan yang dilakukan untuk menguji tingkat keakuratan pada sistem confusion matriks seperti pada gambar dibawah ini.

```
Title : Program pengujian
        keakuratan Gradient Boots
        Classfier pada sistem
        confusion matrix
Implementation
1.  Labels = ['Jawa', 'Nias',
    'Tora']
2.  Cm_Model_Gb =
    Confusion_Matrix(Y_Test,
    Y_Pred, Labels)
3.  Fig =
    Plt.Figure(Figsize=(9,9))
4.  Ax = Fig.Add_Subplot(111)
5.  Sns.Heatmap(Cm_Model_Gb, Annot=True,
    Fmt=".3f", Linewidths=.5, Square =
    True, Cmap= 'Blues_R')
6.  Plt.Ylabel('Actual')
7.  Plt.Xlabel('Predicted')
8.  Ax.Set_Xticklabels(Labels)
9.  Ax.Set_Yticklabels(Labels)
10. Title = 'Gradient Boost Model
    Accuracy Score = '+
    Str(Round(Accuracy_Score_Gb*100,2)) +
    "%"
11. Plt.Title(Title, Size = 15)
```

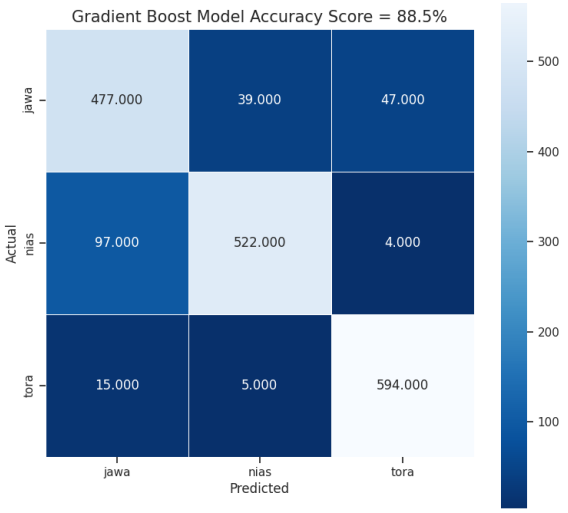
Kode Program 1. *Psuedcode Gradient Boost Classifier*

Perhitungan *Gradient Boost Classifier code* dengan *accuracy gradient boots* dikalikan 100,2, sehingga menghasilkan *Gradient Boots Model Accuracy Score*.

Hasil perhitungan *confusion matrix* dari *Gradient Boots* pada tabel 1 dan *heatmap gradient boots* pada gambar 2.

Tabel 1. *Gradient Boots Result*

	Precision	***Recall***	***Fl-score***	***Support***
Jawa	0.81	0.85	0.83	563
Nias	0.92	0.84	0.88	623
Tora	0.92	0.97	0.94	614
Accuracy			0.89	1800
Mecro	0.88	0.88	0.88	1800
Weight avg	0.89	0.89	0.88	1800



Gambar 2. *Heatmap Confusion Matrix Gradient Boots*

Berdasarkan data pada gambar 2, bahasa Jawa memiliki nilai *true positif* dengan *actual* dan *predicted* yang sama yaitu jawa sehingga menghasilkan nilai yang tinggi 477.000, sedangkan *fals negative* dengan *actual* Nias memiliki *predicted* yang berbeda yaitu Jawa menghasilkan nilai 97.000 dan *fals negative* pada *actual* Tora dan *predicted* Jawa menghasilkan nilai rendah 15.000. Dari Tabel 1 keseluruhan tingkat perhitungan *confusion matrix* dari *gradient boots* mencakup *precision*, *recall* dan *fl-score* menghasilkan *gradient boots model accuracy score* 0.8850 atau 88.5%.

```
predictions = pipeline_dct.predict(["sikilku tatu","noro saoha","meogslina lako"])
predictions
```

```
array(['jawa', 'nias', 'tora'], dtype=object)
```

Gambar 3. *Identification Gradient Boots*

Identification Gradient Boots untuk mengidentifikasi kalimat yang diinputkan dari setiap bahasa daerah, identifikasi kalimat yang dimasukkan kedalam sistem dapat diterjemahkan sesuai dengan kalimat yang sudah ditentukan. Seperti identifikasi kalimat "sikilku tatu", "naro saoha", "meogslina lako" dengan hasil identifikasi bahasa "jawa", "nias", "tora.

Classifier object random forest dengan melakukan *training sampel* data yang dimiliki dari *pickle file* dan menyimpan model ke dalam file untuk dapat diimpor dan digunakan.

Confusion matriks akan menghitung prediksi yang benar dan salah dengan meringkas hasil prediksi dalam masalah klasifikasi, dimana *random forest classifier code* dengan *accuracy gradient boots*

dikalikan 100,2, sehingga menghasilkan *Random Forest Model Accuracy Score*.

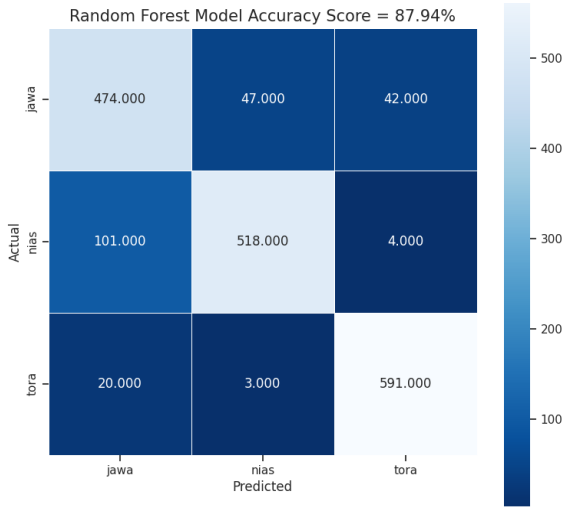
```
Title : Program pengujian
        keakuratan Random Forest
        Classfier pada sistem
        confusion matrix
Implementation
1.  Labels = ['Jawa', 'Nias',
    'Tora']
2.  Cm_Model_Rf =
    Confusion_Matrix(Y_Test,
    Y_Pred, Labels)
3.  Fig =
    Plt.Figure(Figsize=(9,9))
4.  Ax = Fig.Add_Subplot(111)
5.  Sns.Heatmap(Cm_Model_Rf, Annot=True,
    Fmt=".3f", Linewidths=.5, Square =
    True, Cmap = 'Blues_R')
6.  Plt.Ylabel('Actual')
7.  Plt.Xlabel('Predicted')
8.  Ax.Set_Xticklabels(Labels)
9.  Ax.Set_Yticklabels(Labels)
10. Title = 'Random Forest Model
    Accuracy Score = '+
    Str(Round(Accuracy_Score_Rf*100,2))
    + "%"
```

Kode Program 2. *Psuedcode Random Forest Classifier*

Random Forest Result, *accuracy Random Forest Model* = 87.94% dengan perhitungan *confusion matrix* dari *random forest* pada tabel 2 dan *heatmap random forest* pada gambar 4.

Tabel 2. *Random Forest Result*

	Precision	***Recall***	***Fl-score***	***Support***
Jawa	0.80	0.84	0.82	563
Nias	0.91	0.83	0.87	623
Tora	0.93	0.96	0.94	614
Accuracy			0.89	1800
Mecro	0.88	0.88	0.88	1800
Weight avg	0.89	0.88	0.88	1800



Gambar 4. *Heatmap Confusion Matrix Random Forest*

Data pada gambar 4 merupakan hasil *actual* dan *predicted* nilai. Setiap *actual* dan *predicted* yang sama akan menghasilkan nilai yang tinggi atau disebut dengan *true positif* seperti pada *actual* dan *predicted* Jawa menghasilkan nilai 474.000, sedangkan *actual* dan *predicted* berbeda akan menghasilkan nilai yang rendah seperti *actual* Tora dan *predicted* Jawa menghasilkan nilai 20.000. Pada tabel 2 perhitungan *confusion matrix* dari *random forest* mencakup *precision*, *recall* dan *fl-score* menghasilkan *random forest model accuracy score* 0.8794 atau 87.94%.

```
predictions = pipeline_dct.predict(["niha raya","masiang moh","ngelak joe"])
predictions
```

```
array(['nias', 'tora', 'jawa'], dtype=object)
```

Gambar 5. *Identification Random Forest*

Identification Random Forest, mengidentifikasi sistem sesuai dengan bahasa daerah yang diinputkan. Identifikasi kalimat yang dimasukkan kedalam sistem dapat diterjemahkan sesuai dengan kalimat yang sudah ditentukan. Identifikasi kalimat "niha raya", "masiang moh", "ngelak joe" dengan hasil identifikasi bahasa "nias", "tora", "jawa".

Perbandingan nilai akurasi berdasarkan nilai keakuratan pada sistem dapat dilihat pada *confusion matrix* dari metode *gradient boots* dan *random forest* dapat dilihat pada tabel 3.

Tabel 3. Hasil Nilai Akurasi

Model	**Nilai Akurasi**
Gradient Boots	0.8850
Random Forest	0.8794

4. DISKUSI

Pada penelitian sebelumnya membahas mengenai klasifikasi bahasa daerah menggunakan *decision tree* dan *gradient boots*, penelitian ini bertujuan untuk perbandingan kedua metode dalam melakukan klasifikasi bahasa. Dimana penelitian algoritma *naïve bayes* dengan tingkat akurasi sebesar 99,22% dibandingkan dengan *random forest* dengan akurasi 65,44%[6]. Adapun penelitian implementasi metode *Support Vector Machine* (SVM) dan *gradient boots* dalam klasifikasi bahasa daerah (Halmahera, Kalimantan, Toraja), tujuan dari penelitian ini untuk membandingkan kedua metode dalam melakukan klasifikasi bahasa daerah, dimana penelitian menggunakan metode SVM memperoleh hasil sebesar 99,64% dan metode *gradient boots* memperoleh hasil sebesar 65,18%[9]. Pada penelitian ini memiliki tujuan yang sama dengan penelitian sebelumnya yaitu membandingkan kedua metode dalam melakukan klasifikasi bahasa daerah, tetapi dalam penelitian sebelumnya membandingkan antara metode *naïve bayes*, random forest dan *support vector machine*, *gradient boots* sedangkan pada penelitian ini menggunakan perbandingan metode *gradient boots* dan *random forest* dengan hasil algoritma gradient boots yang dapat dilihat pada *Heatmap Confusion true positif* dengan *actual* dan *predicted* yang sama yaitu jawa sehingga menghasilkan nilai yang tinggi 477.000, sedangkan *fals negative* dengan *actual* Nias memiliki *predicted*

yang berbeda yaitu Jawa menghasilkan nilai 97.000, *fals negative* pada *actual* Tora, *predicted* Jawa menghasilkan nilai rendah 15.000 dan perhitungan *confusion matrix* dari *gradient boots* mencakup *precision, recall* dan *fl-score* menghasilkan *gradient boots model accuracy score* sebesar 88,5%, sedangkan algoritma *random forest* memiliki nilai *actual* dan *predicted* yang sama akan menghasilkan nilai yang tinggi atau disebut dengan *true positif* seperti pada *actual* dan *predicted* Jawa menghasilkan nilai 474.000, sedangkan *actual* dan *predicted* berbeda akan menghasilkan nilai yang rendah seperti *actual* Tora, *predicted* Jawa menghasilkan nilai 20.000 dan perhitungan *confusion matrix* dari *random forest* mencakup *precision*, *recall* dan *fl-score* menghasilkan *random forest model accuracy score* menghasilkan nilai sebesar 87,94.

5. KESIMPULAN

Berdasarkan perbandingan yang dilakukan pada penelitian klasifikasi bahasa Jawa, Nias dan Toraja dengan menggunakan metode *gradient boots* dan *random forest*, kedua metode yang digunakan cukup bagus dalam melakukan klasifikasi bahasa maupun penerjemah bahasa dengan hasil tingkat akurasi 0.8 atau 80%. Dengan hasil akhir yang dilihat pada proses perhitungan confusion matrix mencakup nilai *Precision*, *Recall* dan *FI-score* yang menghasilkan nilai *accuracy* pada metode *gradient boots* lebih tinggi dibandingkan *random forest*. Perbandingan kedua metode dapat dilihat pada hasil penelitian *gradient boots* memiliki nilai *accuracy gradient boots* sebesar 0.8850 atau 88.5% dibandingkan dengan metode *random forest* memiliki nilai *accuracy random forest* lebih rendah yaitu 0.8794 atau 87.94%, sehingga pada penelitian ini dapat disimpulkan bahwa metode *gradient boots* memiliki metode yang lebih baik dalam melakuan klasifikasi bahasa. Pada penelitian ini perlu dilanjutkan dengan menggunakan perbandingan metode lain atau meningkatkan metode untuk meningkatkan nilai akurasi.

DAFTAR PUSTAKA

[1] N. Maghfiroh, "Bahasa Indonesia Sebagai Alat Komunikasi Masyarakat Dalam Kehidupan Sehari-Hari," *J. Ilm. Ilmu Komun.*, vol. 19, no. 2, pp. 102–107, 2022.

[2] A. D. Azis, "Bugis Language Maintenance Strategy in Lombok," *J. Pendidik. Bhs. dan Sastra Indones.*, vol. 3, no. 2, pp. 199–208, 2020.

[3] M. Windarti and A. Suradi, "Perbandingan Kinerja 6 Algoritme Klasifikasi Data Mining untuk Prediksi Masa Studi Mahasiswa," *Telematika*, vol. 1, no. 1, pp. 14–30, 2019.

[4] S. Lutfiani, T. H. Saragih, F. Abadi, M. R. Faisal, and K. Dwi, "Perbandingan Metode Extreme Gradient Boosting dan Metode Decision Tree Untuk Klasifikasi Genre Musik," *JIP (Jurnal Inform. Polinema)*, vol. 9, no. 4, pp. 373–382, 2023.

[5] R. Leonardo and J. Pratama, "Perbandingan Metode Random Forest Dan NaïveBayes Dalam Prediksi Keberhasilan Klien Telemarketing," *J. Penelit. Tek. Inform.*, vol. 3, no. 2, pp. 455–459, 2020.

[6] N. Katriani and E. Mailoa, "Klasifikasi Bahasa Daerah Menggunakan Decision Tree Dan Gradient Boots," *J. Tek. Inform. dan Sist. Inf.*, vol. 9, no. 2, pp. 930–940, 2022.

[7] W. Apriliah, I. Kurniawan, M. Baydhowi, and T. Haryati, "Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," *J. Sist. Inf.*, vol. 10, no. 1, pp. 163–171, 2021.

[8] M. Syukron, R. Santoso, and T. Widiharih, "Perbandingan Metode Smote Random Forest Dan Smote Xgboost Untuk Klasifikasi Tingkat Penyakit Hepatitis C Pada Imbalance Class Data," *J. Gaussian*, vol. 9, no. 3, pp. 227–236, 2020.

[9] L. Lumbaa, E. Mailoa, and ..., "Implementasi Metode SVM Dan Gradiant Boost Dalam Klasifikasi Bahasa Daerah (Halmahera, Kalimantan, Toraja)," *J. Tek. Inform. dan Sist. Inf.*, vol. 9, no. 2, pp. 908–915, 2022.

[10] G. A. Mursianto, I. M. Falih, M. Irfan, T. Sakinah, and D. S. Prasvita, "Perbandingan Metode Klasifikasi Random Forest dan XGBoost Serta Implementasi Teknik SMOTE pada Kasus Prediksi Hujan," *J. Senamika*, vol. 2, no. 2, pp. 41–50, 2021.

[11] E. Firasari, U. Khultsum, M. N. Winnarto, and R. Risnandar, "Kombinasi K-NN dan Gradient Boosted Trees untuk Klasifikasi Penerima Program Bantuan Sosial," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 6, pp. 1231--1236, 2020.

[12] Yoga Religia, Agung Nugroho, and Wahyu Hadikristanto, "Klasifikasi Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 187–192, 2021.

[13] E. Renata and M. Ayub, "Penerapan Metode Random forest untuk Analisis Risiko pada dataset Peer to peer lending," *J. Tek. Inform. dan Sist. Inf.*, vol. 6, no. 3, pp. 462–474, 2020.

[14] N. Nuraeni, "Klasifikasi Data Mining untuk Prediksi Potensi Nasabah dalam Membuat Deposito Berjangka," *J. Ilm. Intech Inf. Technol. J. UMUS*, vol. 3, no. 01, pp. 65–75, 2021.

[15] A. C. Nugraha and M. I. Irawan, "Komparasi

Deteksi Kecurangan pada Data Klaim Asuransi Pelayanan Kesehatan Menggunakan Metode Support Vector Machine (SVM) dan Extreme Gradient Boosting (XGBoost)," *J. Sains dan Seni ITS*, vol. 12, no. 1, pp. 40–46, 2023.

[16] S. E. Herni Yulianti, Oni Soesanto, and Yuana Sukmawaty, "Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit," *J. Math. Theory Appl.*, vol. 4, no. 1, pp. 21–26, 2022.

[17] G. M. Momole, "Perbandingan Naïve Bayes dan Random Forest Dalam Klasifikasi Bahasa Daerah," *J. Tek. Inform. dan Sist. Inf.*, vol. 9, no. 2, pp. 855–863, 2022.