JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 1, Maret 2024, Pp. 158-167

ANALISA PERBANDINGAN ALGORITMA RANDOM FOREST DAN NAÏVE BAYES UNTUK KLASIFIKASI CURAH HUJAN BERDASARKAN IKLIM DI INDONESIA

Nicolaus Advendea Prakoso Indaryono*[1)], Rd. Rohmat Saedudin[2)], Faqih Hamami[3)]
1. Telkom University, Indonesia
2. Telkom University, Indonesia
3. Telkom University, Indonesia

Article Info
Kata Kunci: Curah hujan; Data mining; Klasifikasi; *Naïve bayes; Random forest*

Keywords: *Classification; Data mining; Rainfall; Naïve bayes; Random forest*

Article history:
Received 15 November 2023
Revised 29 November 2023
Accepted 13 December 2023
Available online 1 March 2024

DOI :
https://doi.org/10.29100/jipi.v9i1.4421

* Corresponding author.
Nicolaus Advendea Prakoso Indaryono
E-mail address:
nicolausadven@student.telkomuniversity.ac.id

ABSTRAK
Indonesia merupakan wilayah yang hampir seluruhnya memiliki iklim tropis. Sebagai daerah yang kebanyakan memiliki iklim tropis, Indonesia mengalami variasi suhu yang sempit namun variasi curah hujan yang beragam. Curah hujan di Indonesia memiliki tingkat keberagaman yang signifikan. Untuk mencegah bahaya yang ditimbulkan oleh curah hujan tinggi, seperti banjir dan tanah longsor. Selain persiapan terhadap bencana, informasi mengenai curah hujan juga bermanfaat dalam bidang pertanian, transportasi, dan industri. Dengan implementasi klasifikasi data mining, akan membantu proses prediksi curah hujan di Indonesia. Penelitian ini menggunakan data iklim harian di Indonesia dan mengadopsi algoritma Random forest sebagai metode klasifikasi. Algoritma tersebut dipilih karena dapat menghasilkan model klasifikasi yang akurat dan stabil tanpa memerlukan penyesuaian parameter yang kompleks. Selain itu, metode Naïve bayes juga digunakan karena kemudahan implementasinya dan pemodelan probabilitas sederhana yang dapat diterapkan pada berbagai jenis data klasifikasi. Berdasarkan hasil penelitian, mendapatkan kesimpulan bahwa algoritma Random forest mendapatkan performa dan akurasi yang lebih baik daripada algoritma Naïve bayes dalam proses klasifikasi dataset iklim di Indonesia. Algoritma Random forest mencapai akurasi sebesar 86.55%, sementara algoritma Naïve bayes hanya mencapai akurasi sebesar 36.61%. Diharapkan hasil dari penelitian ini dapat digunakan sebagai referensi untuk studi literatur penelitian selanjutnya dan digunakan untuk memantau curah hujan harian di Indonesia guna mencegah bencana alam.

ABSTRACT
Indonesia is a region that almost entirely has a tropical climate. As a mostly tropical climate, Indonesia experiences narrow temperature variations but diverse rainfall variations. Rainfall in Indonesia has a significant degree of variability. To prevent hazards caused by high rainfall, such as floods and landslides. In addition to disaster preparation, rainfall information is also useful in agriculture, transportation, and industry. With the implementation of data mining classification, it will help the rainfall prediction process in Indonesia. This research uses daily climate data in Indonesia and adopts random forest algorithm as a classification method. The algorithm was chosen because it can produce accurate and stable classification models without requiring complex parameter adjustments. In addition, the Naïve Bayes method is also used due to its ease of implementation and simple probability modeling that can be applied to various types of classification data. Based on the research results, it is concluded that random forest algorithm gets better performance and accuracy than Naïve bayes algorithm in the process of climate dataset classification in Indonesia. Random forest algorithm achieved an accuracy of 86.55%, while Naïve bayes algorithm only achieved an accuracy of 36.61%. It is expected that the results of this research can be used as a reference for further research literature studies and used to monitor daily rainfall in Indonesia to prevent natural disasters.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 1, Maret 2024, Pp. 158-167

I. PENDAHULUAN

PENINGKATAN populasi perkotaan yang terus bertumbuh lebih padat atau lebih besar yang mencapai 20 juta penduduk di beberapa kota besar membuat kota-kota sangat rentan terhadap bahaya perubahan iklim. Pemanasan atmosfer menyebabkan meningkatnya degradasi hujan, polusi udara dan juga air [1].

Indonesia merupakan wilayah yang hampir seluruhnya memiliki iklim tropis. Dengan adanya perubahan iklim sendiri berdampak pada perubahan anomali suhu udara rata-rata di Indonesia. Menurut Molle & Larasati, sebagai daerah yang kebanyakan memiliki iklim tropis Indonesia mengalami variasi suhu yang sempit namun variasi curah hujan yang beragam. Perubahan iklim yang terjadi juga akan berpotensi mengakibatkan perubahan curah hujan [2].

Curah hujan adalah jumlah total presipitasi yang terjadi di suatu daerah tertentu, yang dinilai dalam skala harian, mingguan, bulanan, dan tahunan, yang dipengaruhi oleh faktor-faktor tambahan dan merupakan salah satu komponen dari iklim. Di Indonesia, curah hujan memiliki tingkat keberagaman yang signifikan dan merupakan hal yang paling penting dalam kehidupan manusia dan saling terkait dengan kondisi seperti kelembaban, tekanan, suhu, arah serta kecepatan angin [3]. Dengan kondisi intensitas hujan yang tinggi di wilayah Indonesia, diperlukan adanya pengetahuan yang mendukung untuk mencegah bahaya yang ditimbulkan dari bencana akibat curah hujan tinggi seperti banjir dan tanah longsor. Berdasarkan data yang dikumpulkan oleh BNPB, terjadi sebanyak 2.342 kejadian bencana di Indonesia pada tahun 2016. Sebanyak 92% dari kejadian tersebut mayoritasnya adalah bencana hidrometeorologi seperti banjir, longsor, dan puting beliung. Secara khusus, terdapat 766 kejadian banjir, 612 kejadian longsor, dan 74 kejadian gabungan banjir dan longsor. Curah hujan yang tinggi di wilayah terdampak menjadi penyebab utama dari banjir dan tanah longsor tersebut. Tragedi-tragedi ini memiliki dampak serius yang tidak boleh diabaikan. Dampak bencana alam yang terjadi menyebabkan korban jiwa, masalah kesehatan, kerusakan rumah, dan penghancuran gedung-gedung umum. Sebanyak 3,05 juta jiwa mengungsi dan menderita, 522 orang kehilangan nyawa atau dinyatakan meninggal dunia, 69.287 rumah mengalami kerusakan, dan 2.311 bangunan fasilitas umum hancur [4]. Selain melakukan persiapan terhadap bencana yang ditimbulkan, informasi terkait curah hujan juga dapat digunakan untuk berbagai keperluan seperti pada bidang pertanian, transportasi dan industri. Oleh karena itu, diperlukan teknik yang dapat dilakukan untuk melakukan klasifikasi terhadap curah hujan. Salah satu metode yang dapat digunakan untuk melakukan analisa dan pengolahan data agar dapat bermanfaat bagi pembaca ialah metode *data mining*.

Menurut Jackson *Data mining* merupakan proses untuk mengidentifikasi dan menghasilkan suatu data yang berguna dan berkolasi yang dapat dimengerti. Proses keseluruhan dari *data mining* sendiri disebut dengan *Knowledge Discovery in Database (KDD)*. KDD memiliki beberapa tahapan proses berupa persiapan, pemilihan, pembersihan dan interpretasi dari hasil proses *data mining* [5]. Dengan menyaring melalui jumlah data yang telah terkumpul sebelumnya, pertambangan data merupakan proses penemuan koneksi, pola, dan tren baru yang bermanfaat. Dalam rangka mengekstraksi informasi dari kumpulan data besar, disiplin multidisiplin pertambangan data menggabungkan metode dari pembelajaran mesin, pengenalan pola, statistik, basis data, dan visualisasi. [6]. Sedangkan menurut Jollyta dkk, *data mining* adalah metode yang menggunakan kecerdasan buatan, statistik, dan aritmatika untuk mengungkap pengetahuan baru dari sejumlah besar data dalam *database*. *Data mining* adalah teknik yang dapat menghubungkan permintaan konsumen dengan kebutuhan data [7].

Dalam penerapannya penambangan data memiliki beberapa model algoritma yang dapat diimplementasikan diantaranya algoritma *random forest dan naïve bayes*. *Random forest* merupakan sebuah algoritma klasifikasi yang termasuk dalam kategori *ensemble*. Algoritma ini membangun sebuah hutan (*forest*) yang terdiri dari beberapa pohon keputusan (*decision tree*). Dalam melakukan klasifikasi, *Random forest* akan menggunakan teknik *voting* untuk mengambil keputusan dari seluruh pohon keputusan. Dengan menggunakan banyak pohon keputusan, algoritma *random forest* dapat mengatasi masalah yang muncul saat hanya menggunakan satu pohon keputusan dalam klasifikasi. Hal ini dapat meningkatkan nilai akurasi secara keseluruhan dan membuat hasil klasifikasi lebih optimal [8]. Sedangkan meotde *naïve bayes* adalah sebuah metode klasifikasi berdasarkan teknik-teknik peluang dan statistika yang diciptakan oleh seorang ilmuwan asal Inggris bernama Thomas Bayes. Tujuan dari algoritma ini adalah untuk memprediksi hasil yang akan terjadi di masa depan menggunakan hasil masa lalu, sehingga dinamai Teorema Bayes. Teori tersebut kemudian dipadukan dengan kata naïve, yang mengasumsikan situasi di mana setiap atribut memiliki sifat saling bebas [9].

Dalam penelitian [10] yang berjudul Perbandingan Metode Random Forest Dan Naïve Bayes Dalam Prediksi Keberhasilan Klien Telemarketing mengenai prediksi terhadap keputusan klien untuk membantu kinerja telemarketing, mendapatkan hasil perbandingan algoritma naive bayes dan random forest yaitu akurasi naive bayes

sebesar 85% dan random forest 90. Dengan hasil tersebut dapat disimpulkan bahwa pada penelitian yang dilakukan terhadap prediksi keberhasilan klien telemarketing, algoritma random forest lebih baik dalam melakukan klasifikasi tersebut. Selain itu pada penelitian [11] dalam melakukan klasifikasi debitur berdasaran kualitas kredit mendapatkan hasil akurasi sebesar 98.16% pada algoritma random forest dan 95.93 pada algoritma naïve bayes. Pada penelitian terdahulu yang pernah dilakukan, didapatkan hasil bahwa algoritma random forest lebih baik dalam melakukan klasifikasi.

Pada penelitian ini memiliki novelty untuk melakukan klasifikasi curah hujan berdasarkan iklim di Indonesia menggunakan algoritma random forest dan naïve bayes. Selain itu dalam penelitian ini, penulis bertujuan untuk membandingkan kinerja algoritma Random Forest dan Naïve Bayes dalam klasifikasi curah hujan berdasarkan data iklim di berbagai wilayah di Indonesia serta melakukan analisa perbandingan terhadap penggunaan metode algoritma *random forest dan naïve bayes* dengan memanfaatkan data mengenai iklim harian di Indonesia untuk melakukan klasifikasi terhadap curah hujan.

II. METODE PENELITIAN

2.1 Data Mining

Data mining adalah praktik menganalisis secara otomatis dan mengekstraksi pengetahuan dari data menggunakan satu atau lebih algoritma pembelajaran mesin. Proses interaktif dan iteratif ini mencari pola atau model baru yang relevan, praktis, dan dapat dimengerti dalam database besar. Dalam data mining, dilakukan pencarian tren atau pola yang diinginkan dalam dataset yang besar untuk membantu pengambilan keputusan di masa depan. Alat-alat khusus yang memberikan analisis data yang bermakna dan berwawasan dapat mengidentifikasi pola-pola ini, yang kemudian dapat diselidiki lebih lanjut dengan bantuan alat bantu pengambilan keputusan lainnya, jika diperlukan. [12].

2.2 Metode Klasifikasi

Klasifikasi merupakan salah satu pendekatan yang untuk penerapan *data mining*. Metode ini adalah salah satu metode analisis data penghasil model yang mengidentifikasi jenis data utama yang signifikan. Model tersebut, yang disebut sebagai klasifier, dapat digunakan untuk memprediksi keanggotaan suatu data ke dalam kelas yang telah ditentukan [13].

2.3 Random Forest

Random forest merupakan algoritma yang dikembangkan oleh Leo Breimean. *Random forest* adalah sekelompok pohon regresi atau klasifikasi yang tidak dipangkas dan dibuat dengan cara memilih sampel acak dari data. Prediksi dibuat dengan menggabungkan hasil prediksi dari seluruh kelompok pohon regresi atau klasifikasi. *Random forest* memiliki keunggulan seperti mampu mendeteksi kesalahan yang relatif besar, kinerja klasifikasi yang baik, mampu menangani data dengan jumlah sampel yang sedikit, dan metode yang efektif untuk mengestimasi data yang hilang [14]

2.4 Naïve Bayes

Teknik kategorisasi probabilitas *Naïve bayes* menentukan probabilitas total dengan menaikkan frekuensi dan memperhitungkan informasi yang tersedia. Algoritma berbasis konsep Bayes menunjukkan bagaimana semua kualitas ditentukan oleh variabel kelas dan tidak independen atau tidak terkait satu sama lain. Naive bayes dapat digunakan karena menggunakan sejumlah kecil data pelatihan untuk memprediksi parameter terkait klasifikasi dengan andal. Dalam kondisi dunia nyata yang lebih rumit dari yang diantisipasi, teluk naif sering mengungguli prediksi [15].

2.5 SMOTE *(Synthetic Minority Oversampling Technique)*

Pendekatan SMOTE menggunakan konsep oversampling, yang melibatkan penambahan data kelas yang tidak seimbang atau kurang untuk menyeimbangkan angka dengan data dari kelas utama. Metode ini akan menggunakan teknik ketetanggaan untuk membangkitkan data dari kelas minor [16].

2.6 Confusion Matrix

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 1, Maret 2024, Pp. 158-167

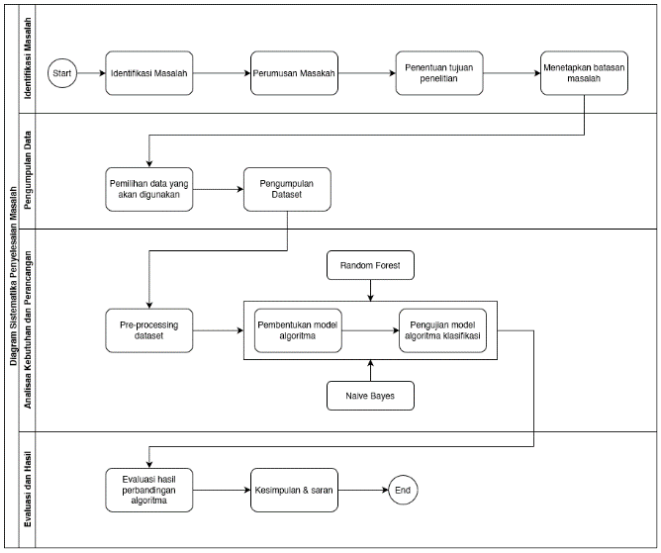
Penambangan data menggunakan *Confusion Matrix* untuk menentukan akurasi prediksi model. Matriks, yang digunakan untuk menentukan nilai akurasi, presisi, *recall,* dan *F1-Score*, terdiri dari data tentang jumlah prediksi yang akurat atau tidak akurat. Fraksi situasi positif yang diproyeksikan benar berdasarkan data aktual disebut sebagai presisi. Persentase kejadian aktual yang secara akurat diantisipasi menjadi positif disebut sebagai *recall*, juga dikenal sebagai sensitivitas. [17]. *Confusion matrix* dapat dilihat pada tabel II.

TABEL I
CONFUSION MATRIX

		Kelas Aktual	
		+	-
Kelas Prediksi	+	True Positives (TP)	False Negatives (FN)
	-	False Positives (FP)	True Negatives (TN)

2.7 Sistematika Penyelesaian Masalah

Terdapat beberapa tahapan dalam proses penyelesaian yang digunakan yaitu, identifikasi masalah, pemilihan data, pemrosesan data, *data mining* dan evaluasi. Sistematika penyelesaian masalah dapat dilihat pada gambar.



Gambar. 1. Sistematika Penyelesaian Masalah

2.8 Pengumpulan Data

Pada penelitian yang dilakukan peneliti untuk membandingkan tingkat akurasi algoritma *random forest dan naïve bayes* adalah data mengenai iklim di Indonesia. Data penelitian iklim di Indonesia sendiri diambil dari situs yang memberikan dan menyediakan data yaitu *Kaggle* dengan alamat : *https://www.kaggle.com/datasets/greegtitan/indonesia-climate/* dan data yang digunakan adalah data dalam kurun waktu 1 tahun terakhir yaitu 2020.

2.9 Identifikasi Data

Pada tahap identifikasi data, dilakukan pemahaman dan analisis terhadap dataset yang telah diperoleh untuk melakukan penyaringan dan pemilihan atribut yang relevan. Atribut yang terpilih akan dilibatkan dalam proses klasifikasi. Pengambilan atribut berdasarkan dari faktor yang dapat mempengaruhi curah hujan. Setelah melakukan proses identifikasi data, proses selanjutnya yaitu menentukan variabel yang akan digunakan untuk penelitian dan membagi dua jenis variabel yaitu variabel independen dan variabel dependen yang dapat dilihat pada tabel II.

TABEL II
VARIABEL PENELITIAN

Variabel	Nama Variabel	Definisi
Variabel In-dependen(X)	*Tavg*	Rata-rata temperatur(°C)
	RH_avg	Rata-rata kelembaban(%)
	ss	Durasi sinar matahari(*hour*)
	Ddd_x	Arah angin dengan kecepatan maksimum(deg)
Variabel De-penden(Y)	*RR*	Curah hujan (mm) Dengan kategori : 0 = ringan 1 = sedang 2 = lebat 3 = sangat lebat

2.10 Data *Preprocessing*

Setelah menetukan atribut yang digunakan, dilakukan proses pencarian data kosong. Hal tersebut diatasi dengan mencari nilai interpolasi pada atribut yang memiliki nilai null yang dapat dilihat pada tabel 3 Penggantian nilai null dengan nilai interpolasi dikarenakan dataset memiliki nilai yang berubah-ubah mengikuti pola seiring berjalannya waktu, oleh karena itu metode penggantian nilai null dengan nilai interpolasi dapat dijadikan pilihan yang baik dengan menggantikan nilai kosong dengan tren data iklim yang ada. Data yang sudah diganti dapat dilihat pada tabel 4.

TABEL III
DATA SEBELUM CLEANSING

Tavg	RH_avg	RR	ss	ddd_x
29.2	74.0	0.0	1.4	280
28.1	78.0	NaN	3.0	260
28.4	81.0	NaN	6.5	260
28.4	80.0	0.0	2.4	260
26.7	86.0	26.6	5.8	350

TABEL IV
DATA SETELAH CLEANSING

Tavg	RH_avg	RR	ss	ddd_x
29.2	74.0	0.0	1.4	280
28.1	78.0	0.0	3.0	260
28.4	81.0	0.0	6.5	260
28.4	80.0	0.0	2.4	260
26.7	86.0	26.6	5.8	350

Setelah tidak ada data kosong, dilakukan proses pemberian label pada dataset yang akan digunakan. Atribut data yang akan diberikan label pada dataset untuk melakukan proses klasifikasi adalah kategori curah hujan berdasarkan probabilistik curah hujan oleh BMKG.

TABEL V
KATEGORI HUJAN

Rentang	Label	Kategori Curah Hujan
0.5 – 20	0	Ringan
20 – 50	1	Sedang
50 – 100	2	Lebat
>100	3	Sangat Lebat

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 1, Maret 2024, Pp. 158-167

Setelah dilakukan proses kategorisasi pada dataset, untuk data yang sudah diberi label dapat dilihat pada tabel 6.

TABEL VI
DATA SETELAH DIBERI LABEL

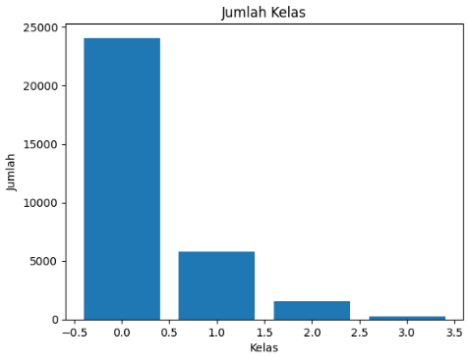
Tavg	RH_avg	ss	ddd_x	label	kategori
27.2	85.0	8.0	180.0	1.0	Ringan
27.2	86.0	6.0	80.0	1.0	Ringan
26.5	88.0	8.5	130.0	1.0	Ringan
27.2	88.0	1.0	130.0	1.0	Ringan
26.9	88.0	3.0	100.0	1.0	Ringan

Berdasarkan hasil data labeling, didapatkan jumlah label pada setiap kelasnya sebagai berikut.

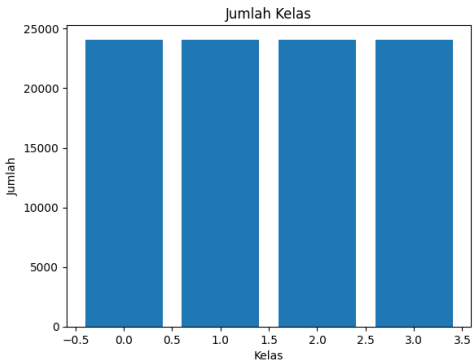
TABEL VII
JUMLAH PEMBAGIAN KELAS

Kelas	Jumlah
0	24.089
1	5.831
2	1.577
3	257
Total	31.754

Setelah mendpatkan hasil labeling data, didapatkan jumlah yang tidak seimbang pada setiap kelasnya yang dapat dilihat pada gambar 2. Untuk mengatasi jumlah data pada setiap kelas yang tidak seimbang, digunakan metode SMOTE untuk melakukan oversampling atau menambah data sintesis terhadap kelas yang tidak seimbang pada kelas yang tidak seimbang pada dataset iklim di Indonesia. Hasil dari pembagian data dengan metode SMOTE.



Gambar. 2. Pembagian kelas sebelum metode SMOTE



Gambar. 3. Pembagian kelas setelah metode SMOTE

2.11 Pemodelan dan Evaluasi

Pada tahap ini dilakukan proses klasifikasi yang dilakukan dengan menerapkan algoritma yang digunakan yaitu naïve bayes dan random forest. Tahap pengujian dibagi menjadi 2 bagian data yaitu data training dan data testing dengan rasio pembagian 80:20. Alat yang digunakan untuk melakukan klasifikasi adalah Google Colab dengan menerapkan bahasa pemrograman python. Setelah dilakukan tahap pemodelan dari kedua algoritma, selanjutnya dilakukan proses evaluasi mengenai perbandingan algoritma dengan metode Confusion Matrix. Tahap evaluasi juga akan menghasilkan nilai akurasi, F1-Score, presisi, dan recall serta menampilkan kurva ROC dan AUC yang optimal dari setiap algoritma klasifikasi yang diuji.

2.12 Normalisasi *MinMax Scaler*

Proses MinMax Scaler bertujuan untuk mengurangi proses komputasi yang diperlukan. Melalui pendekatan ini diharapkan akurasi akan meningkat dan waktu yang dibutuhkan dalam proses pelatihan akan berkurang. Min-Max digunakan untuk mengubah rentang nilai setiap variabel sehingga nilainya berada antara 0 dan 1.

III. HASIL DAN PEMBAHASAN

3.1 Pengujian dengan Naïve Bayes

Pada pengujian dataset dengan algoritma Naïve Bayes, digunakan fungsi dari Gaussian Naïve bayes yang dilakukan dengan import library sklearn pada python. Dengan source code yang dapat dilihat pada gambar 4.

```
model = GaussianNB()
model.fit(X_train, y_train)
```

Gambar. 4. Model naive bayes

Pada tahapan pengujian awal algoritma Naïve bayes untuk mendapatkan hasil akurasi berdasarkan data yang telah dengan rasio 80:20, didapatkan hasil akurasi sebesar 36.61%. Hasil yang didapatkan dari proses evaluasi menggunaka Confusion matrix mendapatkan hasil dengan klasifikasi multi kelas karena terdapat kelas dengan rentang 0-3, tabel confusion matrix dapat dilihat pada tabel 8.

TABEL VIII
CONFUSION MATRIX NAIVE BAYES

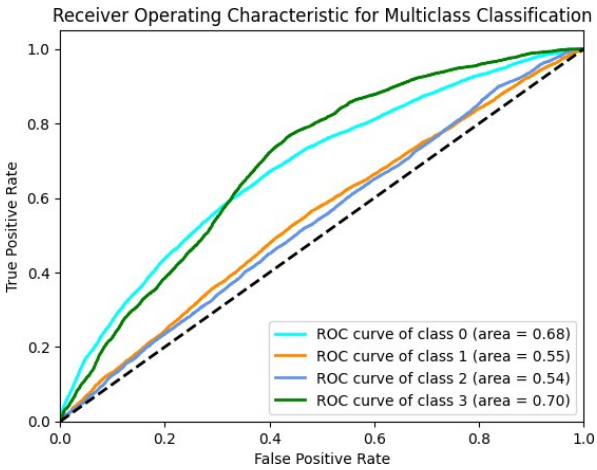
	Pred 1	**Pred 2**	**Pred 3**	**Pred 4**
Actual 0	6445	1037	865	3633
Actual 1	4705	1104	965	5287
Actual 2	3774	989	945	6440
Actual 3	1504	856	483	9146

Dengan hasil evaluasi yang didapatkan dari perhitungan Confusion Matrix, kemudian dapat melakukan proses penghitungan terhadap nilai Recall, Presisi, F1-Score dan skor AUC. Rata-rata nilai yang didapatkn dari setiap perhitungan dapat dilihat pada tabel 11.

TABEL IX
HASIL EVALUASI NAIVE BAYES

Accuracy	36.61%
Precision	33.25%
Recall	36.75%
F1-Score	30.25%
AUC Score	61.75%

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 1, Maret 2024, Pp. 158-167



Gambar. 5. Kurva AUC naive bayes

Grafik dari ROC dan skor dari AUC yang didapatkan dari algoritma Naïve bayes dapat dilihat pada gambar 5.

3.2 Implementasi dengan Random Forest

Pada pengujian dataset dengan algoritma Random Forest digunakan fungsi dari random forest yang dilakukan dengan *import library* sklearn pada python. Dengan *source code* yang dapat dilihat pada gambar 6 berikut.

```
model = RandomForestClassifier(n_estimators=100)
clf = model.fit(X_train, y_train)
```

Gambar. 6. Model random forest

Pada tahapan pengujian awal algoritma Random Forest untuk mendapatkan hasil akurasi berdasarkan data yang telah dengan rasio 80:20, didapatkan hasil akurasi sebesar 86.55%. Hasil yang didapatkan dari proses evaluasi menggunaka Confusion matrix mendapatkan hasil dengan klasifikasi multi kelas karena terdapat kelas dengan rentang 0-3, tabel confusion matrix dapat dilihat pada tabel 10.

TABEL X
CONFUSION MATRIX RANDOM FOREST

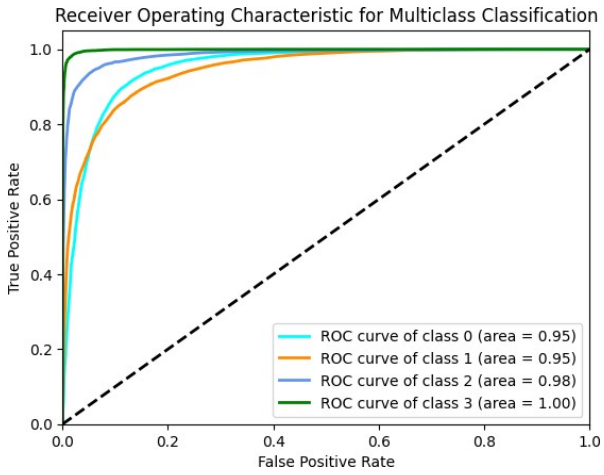
	Pred 1	**Pred 2**	**Pred 3**	**Pred 4**
Actual 0	3876	665	224	**86**
Actual 1	756	3714	235	**100**
Actual 2	189	194	4466	**66**
Actual 3	**31**	**25**	**21**	**4624**

Dengan hasil evaluasi yang didapatkan dari perhitungan Confusion Matrix, kemudian dapat melakukan proses penghitungan terhadap nilai Recall, Presisi, F1-Score dan skor AUC. Rata-rata nilai yang didapatkn dari setiap perhitungan dapat dilihat pada tabel 11.

TABEL XI
HASIL EVALUASI RANDOM FOREST

Accuracy	86.55%
Precision	86.25%
Recall	86.25%
F1-Score	86.5%
AUC Score	97%

Grafik dari ROC dan skor dari AUC yang didapatkan dari algoritma Random Forest dapat dilihat pada gambar 7.



Gambar. 7. Hasil evaluasi random forest

3.3 Evaluasi Perbandingan Algoritma

Hasil perbandingan dari algoritma yang digunakan mendapatkan hasil sebagai berikut.

TABEL XI
PERBANDINGAN ALGORTIMA

	Algoritma	
	Naïve Bayes	***Random forest***
Akurasi Confusion Matrix	36.61%	86.55%
Precision	33.25%	86.25%
Recall	36.75%	86.25%
F1-Score	30.25%	86.5%
AUC Score	61.75%	97%

Berdasarkan perbandingan pada tabel 12, didapatkan hasil dari algoritma naïve bayes dengan nilai akurasi sebesar 36.61%, precision dengan rata-rata 33.25%, recall dengan rata-rata 36.75%, F1-Score dengan nilai rata-rata 30.25% dan AUC Score dengan rata-rata 61.75% dan masuk dalam kategori skor AUC dengan klasifikasi sangat buruk. Kemudian pada algoritma random forest didapatkan nilai akurasi sebesar 86.55%, precision dengan rata-rata 86.25%, recall dengan rata-rata 86.25%, F1-Score dengan nilai rata-rata 86.5% dan AUC Score dengan rata-rata 97% dan masuk dalam kategori skor AUC dengan klasifikasi sangat baik. Berdasarkan hasil tersebut masing-masing algoritma random forest memiliki nilai akurasi yang lebih baik dibandingkan dengan algoritma naïve bayes dalam perbandingan akurasi, presisi, *recall*, *F1-Score* dan skor AUC. Hasil dari perbandingan pada penelitian ini juga menunjukkan bahwa algoritma random forest memiliki kinerja lebih baik dalam melakukan klasifikasi dibandingkan dengan algoritma naïve bayes sesuai dengan penelitian terdahulu yang pernah dilakukan.

IV. KESIMPULAN

Hasil evaluasi perbandinga algoritma Naïve Bayes dan Random Forest, didapatkan hasil dari algoritma Naïve bayes dengan nilai akurasi sebesar 36.61%, Precision dengan rata-rata 33.25%, Recall dengan rata-rata 36.75%, F1-Score dengan nilai rata-rata 30.25% dan AUC Score dengan rata-rata 61.75% dan masuk dalam kategori skor AUC dengan klasifikasi sangat buruk. Kemudian pada algoritma random forest didapatkan nilai akurasi sebesar 86.55%, Precision dengan rata-rata 86.25%, Recall dengan rata-rata 86.25%, F1-Score dengan nilai rata-rata 86.5% dan AUC Score dengan rata-rata 97% dan masuk dalam kategori skor AUC dengan klasifikasi sangat baik. Berdasarkan hasil tersebut masing-masing algoritma random forest memiliki nilai akurasi yang lebih baik dibandingkan dengan algoritma naïve bayes dalam perbandingan akurasi, presisi, recall, F1-Score dan skor AUC.

DAFTAR PUSTAKA

[1] M. M. Rojas-Downing, A. P. Nejadhashemi, T. Harrigan, and S. A. Woznicki, “Climate change and livestock: Impacts, adaptation, and mitigation,” *Clim Risk Manag*, vol. 16, pp. 145–163, 2017, doi: 10.1016/j.crm.2017.02.001.

[2] B. A. Molle and A. F. Larasati, "ANALISIS ANOMALI POLA CURAH HUJAN BULANAN TAHUN 2019 TERHADAP NORMAL CURAH HUJAN (30 TAHUN) DI KOTA MANADO DAN SEKITARNYA," 2020. [Online]. Available: https://web.meteo.bmkg.go.id/id
[3] R. Ruqoyah, Y. Ruhiat, and A. Saefullah, "Analisis Klasifikasi Tipe Iklim Dari Data Curah Hujan Menggunakan Metode Schmidt-Ferguson (Studi Kasus: Kabupaten Tangerang)," 2023.
[4] D. Setiawan, "Analisis Curah Hujan di Indonesia untuk Memetakan Daerah Potensi Banjir dan Tanah Longsor dengan Metode Cluster Fuzzy C-Means dan Singular Value Decompotition (SVD)," *Engineering, MAthematics and Computer Science (EMACS) Journal*, vol. 3, no. 3, pp. 115–120, Oct. 2021, doi: 10.21512/emacsjournal.v3i3.7428.
[5] J. Jackson, "Data Mining; A Conceptual Overview," *Communications of the Association for Information Systems*, vol. 8, 2002, doi: 10.17705/1CAIS.00819.
[6] D. T. Larose and C. D. Larose, "DISCOVERING KNOWLEDGE IN DATA An Introduction to Data Mining Second Edition Wiley Series on Methods and Applications in Data Mining."
[7] Deny Jollyta, William Ramdhan, and Muhammad Zarlis, *Konsep data mining dan penerapan*. Deepublish, 2020.
[8] A. I. Kusumarini, P. A. Hogantara, and N. Chamidah, *Perbandingan Algoritma Random Forest, Naïve Bayes, Dan Decision Tree Dengan Over-sampling Untuk Klasifikasi Bakteri E. Coli*. 2021.
[9] M. F. Rifai, H. Jatnika, and B. Valentino, "Penerapan Algoritma Naïve Bayes Pada Sistem Prediksi Tingkat Kelulusan Peserta Sertifikasi Microsoft Office Specialist (MOS)," *PETIR*, vol. 12, no. 2, pp. 131–144, Sep. 2019, doi: 10.33322/petir.v12i2.471.
[10] R. Leonardo and J. Pratama, "Address: Universitas Prima Indonesia, Teknik Informatika," *Jl. Sekip Sei Kambing Medan*, no. 123, 2011, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Bank+Ma
[11] Bonggo Bawono and Rochdi Wasono, "PERBANDINGAN METODE RANDOM FOREST DAN NAÏVE BAYES UNTUK KLASIFIKASI DEBITUR BERDASARKAN KUALITAS KREDIT," *Seminar Nasional Edusaintek*, 2022.
[12] Erma Delima Sikumbang, "Penerapan Data Mining Penjualan Sepatu Menggunakan Metode Algoritma Apriori," *Jurnal Teknik Komputer*, vol. 4, no. 1, 2018, doi: https://doi.org/10.31294/jtk.v4i1.2560.
[13] J. Han, M. Kamber, and J. Pei, *Data Mining*. Elsevier, 2012. doi: 10.1016/C2009-0-61819-5.
[14] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random Forests and Decision Trees," 2012. [Online]. Available: www.IJCSI.org
[15] A. Triawan and D. Lintang Melinda, "Penerapan Metode Naïve Bayes Untuk Rekomendasi Topik Tugas Akhir Berdasarkan Daftar Hasil Studi Mahasiswa di Perguruan Tinggi," *Teknois : Jurnal Ilmiah Teknologi Informasi dan Sains*, vol. 10, no. 2, pp. 58–70, Nov. 2020, doi: 10.36350/jbs.v10i2.91.
[16] 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", [Online]. Available: https://ejournal3.undip.ac.id/index.php/gaussian/
[17] G. A. Sandag, "Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest," *CogITo Smart Journal*, vol. 6, no. 2, pp. 167–178, Dec. 2020, doi: 10.31154/cogito.v6i2.270.167-178.