Jurnal Informatika dan Teknologi Interaktif
Volume 2, Nomor 2, Agustus 2025

DOI : 10.63547/jiite.v2i2.68
e-ISSN : 3048-1023

Optimasi Prediksi Bencana Banjir Menggunakan Teknik Smote Berbasis Algortima Naive Bayes

Refida Septiana Putri[1], Reykha Putri Randika[2], Atiqah Noor Zhaafirah[3], Febi Dwi Sasmita[4], Adhisa Nanda Kurnia[5], Albab Muzaki [6]

[1,2,3,4,5]Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Amikom Purwokerto
[6]Program Studi Teknik Informatika, Fakultas Teknik dan Sains, Universitas Muhammadiyah Purwokerto
surel:[1]refidaseptianaputri0@gmail.com,[2]reykhaputri690@gmail.com,[3]viraira99@gmail.com,[4]dwisasmitssssafebi@gmail.com, [5]adhisanandak@gmail.com,[6]albabmuzaki31@gmail.com,

Info Artikel

Sejarah artikel:
Diterima 11-06-2025
Revisi 09-07-2025
Diterima 19-07-2025

Kata kunci:
Banjir
Naive Bayes
SMOTE
CRISP-DM

ABSTRAK

Banjir merupakan salah satu bencana alam yang memberikan dampak signifikan terhadap kehidupan sosial, ekonomi, dan infrastruktur. Peningkatan intensitas curah hujan akibat perubahan iklim global serta sistem drainase yang tidak memadai menjadi faktor utama penyebab banjir di berbagai wilayah. Oleh karena itu, kemampuan untuk memprediksi banjir secara dini menjadi hal yang krusial dalam upaya mitigasi risiko. Penelitian ini mengembangkan model klasifikasi banjir berbasis algoritma *Naive Bayes* yang dipadukan dengan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) untuk mengatasi ketidakseimbangan data antara kelas banjir dan non-banjir. Proses analisis dilakukan dengan pendekatan CRISP-DM yang mencakup enam tahapan: pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan implementasi. Dataset yang digunakan merupakan data historis curah hujan wilayah Kerala, India, periode 1901–2018 yang diperoleh dari platform Kaggle. Evaluasi performa model dilakukan menggunakan metrik Accuracy, Precision, Recall, F1-Score, dan Hamming Loss. Hasil penelitian menunjukkan bahwa kombinasi *Naive Bayes* dan SMOTE mampu meningkatkan sensitivitas model terhadap kejadian banjir, serta memberikan kontribusi dalam pengembangan sistem peringatan dini bencana yang lebih efektif dan andal.

Penulis yang sesuai:
Refida Septiana Putri
Program Studi Informatika Fakultas Ilmu Komputer Universitas Amikom Purwokerto
Email: refidaseptianaputri0@gmail.com

1. PENDAHULUAN

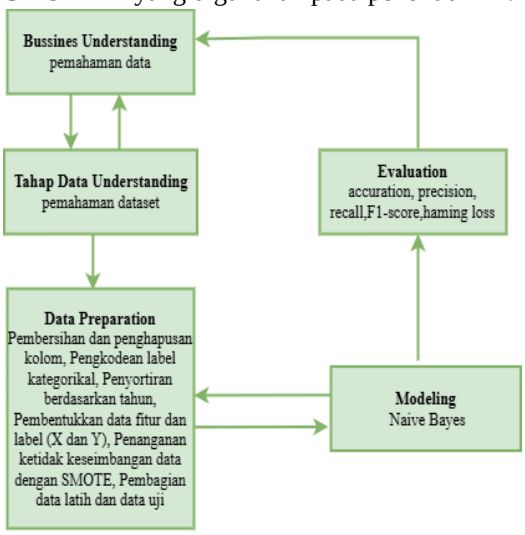
Banjir merupakan salah satu bencana alam yang memiliki dampak signifikan terhadap kehidupan sosial, ekonomi, dan infrastruktur [1]. Faktor terjadinya banjir bisa dari curah hujan ekstrim dan sistem drainase yang memadai [2] Peningkatan intensitas curah hujan akibat perubahan iklim global memperbesar risiko banjir di berbagai wilayah. Oleh karena itu, kemampuan untuk memprediksi potensi terjadinya banjir secara dini sangat penting guna mendukung mitigasi risiko dan pengambilan keputusan yang cepat dan tepat oleh pihak berwenang. Dalam beberapa

tahun terakhir, pendekatan berbasis data dan kecerdasan buatan (AI) telah dimanfaatkan secara luas untuk melakukan prediksi bencana, termasuk banjir. Salah satu metode yang banyak digunakan adalah algoritma Naive Bayes karena kemudahannya dalam implementasi dan efisiensinya dalam klasifikasi[3]. Namun, tantangan utama dalam pemodelan prediksi banjir adalah adanya ketidakseimbangan data (imbalanced dataset), di mana jumlah kejadian banjir (kelas minoritas) jauh lebih sedikit dibandingkan dengan data non-banjir (kelas mayoritas). Ketidakseimbangan ini dapat menyebabkan penurunan kinerja model dalam mengenali kejadian banjir secara akurat. Dalam bidang prediksi banjir menunjukkan bahwa algoritma Naive Bayes mampu mengklasifikasikan kejadian banjir dengan tingkat akurasi mencapai 79,16%, berdasarkan data historis curah hujan[4]. Algoritma ini terbukti cukup efektif dalam memodelkan hubungan antara intensitas curah hujan dan kemungkinan terjadinya banjir, meskipun tetap menghadapi tantangan dalam mendeteksi kasus-kasus yang jarang terjadi atau termasuk dalam kelas minoritas. Meskipun demikian, tantangan utama dalam model prediksi banjir adalah ketidakseimbangan data (imbalanced dataset), di mana jumlah data banjir sebagai kelas minoritas lebih sedikit dibandingkan kelas mayoritas. Oleh karena itu, teknik *Synthetic Minority Over-sampling Technique* (SMOTE) digunakan untuk menyeimbangkan data dengan menciptakan sampel sintetis pada kelas minoritas, sehingga meningkatkan sensitivitas model terhadap kejadian banjir yang jarang terjadi namun krusial [5].

Dalam penelitian ini, proses analisis data bencana banjir dilakukan menggunakan pendekatan CRISP-DM (Cross Industry Standard Process for Data Mining) yang terdiri dari enam tahapan utama: pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan implementasi [6]. Pendekatan ini belum banyak diterapkan secara komprehensif dalam penelitian kebencanaan di Indonesia, yang umumnya langsung fokus pada performa algoritma [7]. Dengan mengadopsi CRISP-DM, penelitian ini bertujuan mengembangkan model klasifikasi banjir berbasis algoritma Naive Bayes dan SMOTE dengan data historis curah hujan wilayah Kerala, India dari tahun 1901–2018 yang diperoleh dari Kaggle. Evaluasi model dilakukan menggunakan metrik Accuracy, Precision, Recall, F1-Score, dan Hamming Loss untuk memastikan keandalan model dalam mendeteksi banjir secara dini. Hasil yang diperoleh diharapkan dapat meningkatkan efektivitas sistem peringatan dini bencana dan pengambilan keputusan oleh pemangku kebijakan[8].

2. METODE

Dalam penelitian ini metode yang diterapkan yaitu CRISP-DM (Cross-Industry Standard Process for Data Mining) sebuah metode standar dalam proses pengembangan sistem berbasis data mining yang mempunyai 6 Tahapan: yaitu business understanding, tahap data understanding, tahap data preparation, tahap modeling dan tahap evaluation. Dan pada gambar 1 adalah alur proses CRISP-DM yang digunakan pada penelitian ini.



Gambar 1. Proses alur penelitian berdasarkan CRISP-DM

2.1. Business Understanding

Prediksi banjir seringkali menghadapi tantangan karena data kejadian banjir yang tidak seimbang, di mana jumlah kasus banjir (kelas minoritas) jauh lebih sedikit dibandingkan dengan kondisi normal (kelas mayoritas). Hal ini dapat menyebabkan rendahnya akurasi model dalam mendeteksi potensi banjir secara tepat. Permasalahan Dataset tidak seimbang (imbalanced), dimana kasus banjir (minoritas) jauh lebih sedikit daripada non-banjir (mayoritas). Tanpa penanganan yang tepat, model Naive Bayes cenderung bias ke kelas mayoritas (non-banjir), sehingga gagal memprediksi banjir secara akurat.

2.2. Tahap Data Understanding

Data yang digunakan dalam penelitian ini diperoleh dari situs *Kaggle*, melalui dataset berjudul *Monthly Rainfall Index and Flood Probability* yang disusun oleh Mukul Thakur (2020). Dataset ini berisi data historis curah hujan serta kejadian banjir tahunan di negara bagian Kerala, India, yang tercatat sejak tahun 1901 hingga 2018. Dataset ini terdiri dari 118 entri (baris) dan 16 atribut (kolom) seperti pada tabel dibawah :

Tabel 1. Keterangan dataset *Monthly Rainfall Index and Flood Probability*

No	Nama Kolom	Deskripsi
1	SUBDIVISION	Subdivisi wilayah pencatatan (hampir seluruhnya "KERALA")
2	YEAR	Tahun pencatatan data (1901–2018)
3	JAN	Curah hujan bulan Januari (dalam mm)
4	FEB	Curah hujan bulan Februari(dalam mm)
5	MAR	Curah hujan bulan Maret (dalam mm)
6	APR	Curah hujan bulan April(dalam mm)
7	MAY	Curah hujan bulan Mei(dalam mm)
8	JUN	Curah hujan bulan Juni (dalam mm)
9	JUL	Curah hujan bulan Juli (dalam mm)
10	AUG	Curah hujan bulan Agustus (dalam mm)
11	SEP	Curah hujan bulan September (dalam mm)
12	OCT	Curah hujan bulan Oktober (dalam mm)
13	NOV	Curah hujan bulan November (dalam mm)
14	DEC	Curah hujan bulan Desember (dalam mm)
15	ANNUAL RAINFALL	Total curah hujan tahunan (range: 2.068,8–4.473,0)

Atribut FLOODS digunakan sebagai variabel target dalam penelitian ini dan terdiri dari dua kelas: "YES" menandakan terjadinya banjir, dan "NO" menandakan tidak terjadi banjir yaitu sebanyak 60 data poin mengalami banjir dan 58 data poin tidak. Hampir seluruh entri pada kolom SUBDIVISION bernilai “KERALA”, namun ditemukan satu entri dengan perbedaan penulisan yang dapat dikoreksi pada tahap pra proses data. Pola curah hujan bulanan menunjukkan karakteristik musiman khas wilayah tropis, dengan puncak curah hujan pada bulan Juni hingga Agustus. Total curah hujan tahunan bervariasi dari 2.068,8 mm hingga 4.473,0 mm, yang mencerminkan dinamika iklim dari tahun ke tahun. Seluruh nilai dalam dataset ini tercatat lengkap tanpa adanya nilai kosong (missing values), sehingga dataset ini siap digunakan dalam analisis statistik dan pemodelan prediktif. Dataset ini tersedia secara publik dan dapat diakses https://www.kaggle.com/datasets/mukulthakur177/kerela-flood

2.3. Tahap Data Preparation

Pre-processing data merupakan langkah penting untuk memastikan data siap dianalisis dengan penghapusan ketidakkonsistenan, penanganan data tidak seimbang, serta pengkodean agar data dapat dipahami oleh algoritma. Berikut tahapan preparation yang dilakukan

2.3.1. Pembersihan dan Penghapusan Kolom kosong

Dengan mengidentifikasi identifikasi data yang hilang, penghapusan kolom yang dianggap tidak relevan.bisa meningkatkan kualitas dataset sebelum pemodelan.

2.3.2. Pengkodean Label Kategorikal

Mengkategorikan Data yang perlu dikonversi menjadi format numerik agar dapat diproses oleh algoritma machine learning.

2.3.3. Penyortiran Pembentukan Data Fitur dan Label (x dan y)

Berdasarkan Tahun model dapat dilatih dan diuji dengan data yang terdistribusi secara kronologis, menjaga urutan historis yang penting dalam konteks prediksi fenomena alam seperti banjir. fitur x untuk merepresentasikan data curah hujan bulanan hingga total tahunan dan fitur y untuk menunjukkan apakah terjadi banjir pada tahun tersebut.

2.3.4. Pembagian Data Latih dan Uji

Dataset dibagi menjadi data latih dan data uji menggunakan stratified split dengan rasio 70:30 dengan data yang sedikit Rasio ini dipilih karena dianggap memberikan keseimbangan optimal antara data yang cukup untuk pelatihan model, dan data yang cukup untuk melakukan evaluasi performa.

2.3.5. Penanganan Ketidakseimbangan Data

Pada dataset banjir memiliki distribusi yang tidak seimbang antara kejadian banjir dan non-banjir. Teknik resampling, jadi teknik SMOTE (Synthetic Minority Oversampling Technique) diterapkan untuk menyeimbangkan dataset, sehingga model tidak bias terhadap kelas mayoritas dan dapat mendeteksi kejadian banjir dengan lebih akurat.

2.4. Tahap Modeling

Untuk tahapan modeling ini dilakukan metode SMOTE dengan algoritma Naive Bayes, yang diterapkan pada data banjir negara bagian kerala yang bersumber dari kaggle, teknik pra pemrosesan data dan metode evaluasi dengan perangkat lunak python sebagai pendukung penerapan model dan algoritma yang sudah disebutkan Guna mencapai tujuan penelitian yaitu menghasilkan prediksi banjir dengan optimal.

2.4.1. SMOTE

Synthetic Minority Oversampling Technique (SMOTE) merupakan teknik yang menyeimbangkan dataset kelas minoritas yang mungkin menimbulkan masalah hingga seimbang menjadi data mayoritas[9]. SMOTE (Synthetic Minority Over-sampling Technique) ialah salah satu metode yang banyak digunakan untuk mengatasi ketidakseimbangan kelas dalam dataset. Teknik ini berkerja dengan cara mengambil sampel data baru. Jumlah data sampel yang diambil menyesuaikan dengan jumlah data minoritas[10].

2.4.2. Naive bayes

Algoritma Naïve Bayes merupakan salah satu teknik klasifikasi yang digunakan untuk memprediksi suatu kejadian berdasarkan nilai probabilitas. Algoritma ini mengacu pada Teorema Bayes, yang memungkinkan perhitungan probabilitas suatu hipotesis dengan mempertimbangkan informasi awal (prior) dan data baru yang diperoleh (likelihood). Melalui pendekatan ini, nilai probabilitas akhir (posterior) dapat diperbarui secara lebih akurat berdasarkan bukti terkini, sehingga mendukung proses pengambilan keputusan yang lebih tepat[11].

2.5. Tahap Evaluasi

Evaluasi terhadap performa model sangat penting untuk mengetahui sejauh mana kemampuan model dalam mengenali pola yang tepat dari data yang digunakan.evaluasi ini menggunakan Lima metrik utama, yaitu Accuracy, Precision, Recall, F1 Score dan Hamming Loss yang seluruhnya dihitung menggunakan pendekatan Confusion Matrix.

2.5.1. Accuracy

Accuracy adalah ukuran proporsi dari keseluruhan prediksi yang dilakukan model yang benar. Dalam konteks klasifikasi, accuracy menghitung seberapa sering model membuat prediksi yang tepat, baik untuk kelas positif maupun negatif .

2.5.2. Precision

Precision adalah ukuran yang menunjukkan ketepatan model dalam memprediksi kelas positif. Precision menjawab pertanyaan: “Dari semua prediksi yang dikatakan sebagai positif oleh model, berapa banyak yang benar-benar positif?”. Precision diartikan sebagai rasio item relevan yang dipilih terhadap semua item terpilih. Precision merupakan probabilitas bahwa sebuah item yang dipilih adalah relevan. Dengan kata lain precision diartikan sebagai kecocokan antara permintaan informasi dengan jawaban atas permintaan tersebut [12].

2.5.3. Recall

Recall merupakan jumlah prediksi yang relevan dengan label kebenaran dibandingkan dengan semua label bernilai 1 pada label kebenaran [13]. Recall, atau sensitivitas, mengukur kemampuan model untuk menemukan seluruh kasus yang benar-benar positif. Dalam konteks ini, recall menjawab: “Dari semua kejadian hujan sangat lebat yang sebenarnya terjadi, berapa banyak yang berhasil dikenali oleh model?”.

2.5.4. F1-Score

F1-Score merupakan salah satu metrik evaluasi yang digunakan untuk mengukur kinerja model klasifikasi dengan menyeimbangkan antara Presisi (Precision) dan Sensitivitas (Recall). F1-Score sangat berguna ketika data yang digunakan tidak seimbang atau ketika kedua jenis kesalahan False Positive (positif palsu) dan False Negative (negatif palsu) memiliki dampak yang sama seriusnya [14]. F1-Score adalah rata-rata harmonik dari presisi dan recall. Tidak seperti rata-rata aritmatika biasa, rata-rata harmonik memberikan lebih banyak bobot pada nilai yang lebih kecil, sehingga F1-Score akan rendah jika salah satu dari precision atau recall rendah, sekalipun yang lain tinggi.

2.5.5. Hamming Loss

Hamming Loss adalah salah satu metrik evaluasi yang umum digunakan dalam klasifikasi multi label. Metrik ini menghitung rata-rata kesalahan prediksi label terhadap jumlah total label yang tersedia. Karena hamming loss menyatakan nilai rata-rata suatu label salah diprediksi, maka performa model dikatakan lebih baik ketika memiliki nilai hamming loss yang semakin kecil [15].

3. HASIL DAN PEMBAHASAN

3.1. Data Preparation

Dari metode CRISP-DM yang diterapkan setelah memahami tujuan dan dataset maka dilakukan pra pemrosesan dari dataset dengan mulai pembersihan dataset yaitu dengan penghapusan kolom SUBMISSION karena berisi data yang bernilai string setelah itu cek nilai kosong, dan hasilnya dataset tidak terdapat data yang kosong

```
In [1]: runfile('D:/anaconda3/filepraktikumPM/projek/ini jadi SMOTE ke
data training.py', wdir='D:/anaconda3/filepraktikumPM/projek')
Cek Dataset:
SUBDIVISION          0
YEAR                 0
JAN                  0
FEB                  0
MAR                  0
APR                  0
MAY                  0
JUN                  0
JUL                  0
AUG                  0
SEP                  0
OCT                  0
NOV                  0
DEC                  0
 ANNUAL RAINFALL     0
FLOODS               0
dtype: int64
```

Gambar 2. Hasil dari cek nilai kosong

Karna dalam dataset tidak terdapat nilai kosong selanjutnya pengkodean label kategorikal dimana melakukan perubahan pada kolom FLOODS yang terdapat kata YES dan NO dirubah menjadi 1 dan 0, karna dalam algortima naive bayes memproses data dalam bentuk numerik agar mendapatkan hasil yang akurat, dan hasilnya terdapat pada gambar 3.

Index	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL RAINFALL	FLOODS
0	1901	28.7	44.7	51.6	160	174.7	824.6	743	357.5	197.7	266.9	350.8	48.4	3248.6	1
1	1902	6.7	2.6	57.3	83.9	134.5	390.9	1205	315.8	491.6	358.4	158.3	121.5	3326.6	1
2	1903	3.2	18.6	3.1	83.6	249.7	558.6	1022.5	420.2	341.8	354.1	157	59	3271.2	1
3	1904	23.7	3	32.2	71.5	235.7	1098.2	725.5	351.8	222.7	328.1	33.9	3.3	3129.7	1
4	1905	1.2	22.3	9.4	105.9	263.3	850.2	520.5	293.6	217.2	383.5	74.4	0.2	2741.6	0
5	1906	26.7	7.4	9.9	59.4	160.8	414.9	954.2	442.8	131.2	251.7	163.1	86	2708	0
6	1907	18.8	4.8	55.7	170.8	101.4	770.9	760.4	981.5	225	309.7	219.1	52.8	3671.1	1
7	1908	8	20.8	38.2	102.9	142.6	592.6	902.2	352.9	175.9	253.3	47.9	11	2648.3	0
8	1909	54.1	11.8	61.3	93.8	473.2	704.7	782.3	258	195.4	212.1	171.1	32.3	3050.2	1
9	1910	2.7	25.7	23.3	124.5	148.8	680	484.1	473.8	248.6	356.6	280.4	0.1	2848.6	0

Gambar 3. Hasil proses data FLOODS

Selanjutnya dalam proses penyortiran data dengan pembentukan variabel x independen dari kolom 2 hingga 14 sebagai kolom perbandingan dan y dependen kolom 1 sebagai variabel target prediksi.

```
36
37  # 5. Pengambilan data X dan y dan sort
38  df2 = df2.sort_values('YEAR').reset_index(drop=True)
39  x = df2.iloc[:,1:14]
40  y = df2.iloc[:,-1]
41
```

Gambar 4. Pembentukan data x dan y

Dengan telah dilakukannya pra pemrosesan data maka sebelum dilakukannya SMOTE data perlu dilakukan pembagian data latih dan data uji, dengan menerapkan perbandingan 0.7 : 0.3 yang artinya 70% data latih dan 30% data uji.

```
41
42  # 6. Split 70:30, train dan test
43  X_train, X_test, y_train, y_test = train_test_split(
44      x, y, test_size=0.3, shuffle=True, random_state=42)
45
```

Gambar 5. Pembagiann data latih dan uji

Setelah data dibersihkan, dikodekan, dan dibagi secara proporsional ke dalam data latih dan data uji, teknik SMOTE diterapkan pada data latih untuk menyeimbangkan distribusi kelas antara kejadian banjir dan non-banjir. Teknik SMOTE bertujuan untuk meningkatkan kinerja model klasifikasi dengan memberikan representasi yang lebih adil bagi kelas minoritas selama proses pelatihan model.

```
45
46  # 7. Penerapan SMOTE setelah split
47  from imblearn.over_sampling import SMOTE
48  smote = SMOTE(random_state=42)
49  X_train, y_train = smote.fit_resample(X_train, y_train)
50
```

Gambar 6. Penerapan SMOTE

3.2. Modeling dan Evaluasi algoritma Naive Bayes

Dataset yang digunakan berasal dari file yang berformat CSV, yang mencangkup atribut dari kolom YEARS hingga FLOODS. Dataset tersebut telah melalui tahap preprocessing untuk memastikan kelayakan data sebelum digunakan dalam proses klasifikasi. 10 data teratas dari hasil preprocessing adalah sebagai berikut:

Index	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	NUAL RAINFA	FLOODS
0	1901	28.7	44.7	51.6	160	174.7	824.6	743	357.5	197.7	266.9	350.8	48.4	3248.6	1
1	1902	6.7	2.6	57.3	83.9	134.5	390.9	1205	315.8	491.6	358.4	158.3	121.5	3326.6	1
2	1903	3.2	18.6	3.1	83.6	249.7	558.6	1022.5	420.2	341.8	354.1	157	59	3271.2	1
3	1904	23.7	3	32.2	71.5	235.7	1098.2	725.5	351.8	222.7	328.1	33.9	3.3	3129.7	1
4	1905	1.2	22.3	9.4	105.9	263.3	850.2	520.5	293.6	217.2	383.5	74.4	0.2	2741.6	0
5	1906	26.7	7.4	9.9	59.4	160.8	414.9	954.2	442.8	131.2	251.7	163.1	86	2708	0
6	1907	18.8	4.8	55.7	170.8	101.4	770.9	760.4	981.5	225	309.7	219.1	52.8	3671.1	1
7	1908	8	20.8	38.2	102.9	142.6	592.6	902.2	352.9	175.9	253.3	47.9	11	2648.3	0
8	1909	54.1	11.8	61.3	93.8	473.2	704.7	782.3	258	195.4	212.1	171.1	32.3	3050.2	1
9	1910	2.7	25.7	23.3	124.5	148.8	680	484.1	473.8	248.6	356.6	280.4	0.1	2848.6	0
10	1911	3	4.3	18.2	51	180.6	990	705.3	178.6	60.2	302.3	145.7	87.6	2726.7	0

Gambar 7. Dataset setelah preprocessing

Tahap splitting data dilakukan menggunakan metode train-test split dengan rasio 70:30 seperti yang sudah dijelaskan sebelumnya. Untuk mengatasi ketidakseimbangan kelas variable target, diterapkan metode SMOTE pada data latih. Proses klasifikasi dilakukan menggunakan algoritma Naive Bayes Gaussian. Model yang telah dilatih kemudian diuji terhadap data uji , dan hasil evaluasinya sebagai berikut:

```
Evaluasi Naive Bayes(% persen)
Akurasi       : 91.67%
Presisi       : 90.00%
Recall        : 94.74%
F1-Score      : 92.31%
Hamming Loss  : 8.33%
```

Gambar 8. Evaluasi Naive Bayes

Berdasarkan hasil evaluasi tersebut, model menunjukan performa klasifikasi yang cukup baik, dengan akurasi algoritma Naive Bayes sebesar 91,67% dan nilai F1-Score sebesar 92,31%, yang menunjukan keseimbangan antara presisi dan recall. Nilai hamming loss yang rendah sebesar 8,33% juga menunjukan tingkat kesalahan yang relatif kecil.

4. KESIMPULAN

Penelitian ini berhasil membuktikan bahwa kombinasi algoritma Naive Bayes dan teknik SMOTE mampu meningkatkan akurasi dan sensitivitas dalam prediksi bencana banjir. Permasalahan ketidakseimbangan data yang

menjadi tantangan utama dalam pemodelan prediksi banjir dapat diatasi dengan efektif melalui penerapan SMOTE, yang menyeimbangkan distribusi antara data banjir dan non-banjir. Penerapan proses CRISP-DM juga memberikan alur kerja yang sistematis dan komprehensif, mulai dari pemahaman bisnis hingga tahap evaluasi.

Hasil evaluasi model menunjukkan kinerja yang sangat baik dengan akurasi sebesar 91,67%, F1-Score sebesar 92,31%, dan nilai Hamming Loss yang rendah sebesar 8,33%, yang menandakan model memiliki tingkat kesalahan prediksi yang kecil. Temuan ini menunjukkan bahwa pendekatan yang digunakan tidak hanya andal, tetapi juga layak diterapkan dalam sistem peringatan dini bencana banjir. Kedepan, hasil penelitian ini dapat dikembangkan lebih lanjut dengan menggunakan dataset yang lebih besar dan beragam, serta menggabungkan variabel lingkungan lainnya seperti data topografi dan tata guna lahan untuk meningkatkan akurasi dan generalisasi model. Selain itu, pendekatan ini memiliki potensi untuk diterapkan di wilayah lain dengan karakteristik curah hujan dan banjir yang serupa, guna mendukung mitigasi risiko dan pengambilan keputusan yang lebih cepat dan tepat oleh pemangku kebijakan.

REFERENSI:

[1] D. D. Utomo and F. Y. D. Marta, "Dampak Bencana Alam Terhadap Perekonomian Masyarakat di Kabupaten Tanah Datar," *J. Terap. Pemerintah. Minangkabau*, vol. 2, no. 1, pp. 92–97, 2022, doi: 10.33701/jtpm.v2i1.2395.

[2] A. Kurnia, "Pengembangan Model Prediksi Banjir Urban dengan Menggunakan Data Hujan dan Data Geospasial," *WriteBox*, pp. 1–12, 2024, [Online]. Available: https://writebox.cloud/index.php/wb/article/view/54%0Ahttps://writebox.cloud/index.php/wb/article/download/54/54

[3] S. A. Saputra and H. Soetanto, "Implementasi Naïve Bayes untuk Klasifikasi Prediksi serta Analisis Data Banjir di Wilayah Jakarta Pusat," *Semin. Nas. Mhs. Fak. Teknol. Inf.*, vol. 3, no. September, pp. 296–304, 2024.

[4] D. Ariyadi, T. A. Y. Siswa, and R. Rudiman, "Penerapan Metode PSO-SMOTE Pada Algoritma Random Forest Untuk Mengatasi Class Imbalance Data Bencana Tanah Longsor," *Kesatria J. Penerapan Sist. Inf. (Komputer dan Manajemen)*, vol. 6, no. 1, pp. 320–329, 2025, doi: 10.30645/kesatria.v6i1.574.

[5] R. Nursyahfitri, C. Rozikin, and R. I. Adam, "Penerapan Metode SMOTE dalam Klasifikasi Daerah Rawan Banjir di Karawang Menggunakan Algoritma Naive Bayes," *J. Sist. dan Teknol. Inf.*, vol. 10, no. 4, p. 339, 2022, doi: 10.26418/justin.v10i4.46935.

[6] Y. Suhanda, I. Kurniati, and S. Norma, "Penerapan Metode Crisp-DM Dengan Algoritma K-Means Clustering Untuk Segmentasi Mahasiswa Berdasarkan Kualitas Akademik," *J. Teknol. Inform. dan Komput.*, vol. 6, no. 2, pp. 12–20, 2020, doi: 10.37012/jtik.v6i2.299.

[7] A. R. Susmita, E. Darmawi, and E. W. Suri, "Analisis Pendekatan Mitigasi Bencana Alam Di Badan Nasional Penanggulangan Bencana Kota Bengkulu," *Masy. Demokr. - J. Ilm. Adm. Publik*, vol. 1, no. 2, pp. 9–18, 2023, doi: 10.32663/md.v1i2.3812.

[8] "21.55.1088 Slamet Triyanto.pdf."

[9] V. J. Rivaldo, T. A. Y. Siswa, and W. J. Pranoto, "Perbaikan Akurasi Naïve Bayes dengan Chi-Square dan SMOTE Dalam Mengatasi High Dimensional dan Imbalanced Data Banjir," *J. Media Inform. Budidarma*, vol. 8, no. 3, p. 1656, 2024, doi: 10.30865/mib.v8i3.7886.

[10] Hizbul Izzi, Arief Setyanto, and Anggit Dwi Hartanto, "Optimalisasi Akurasi Algoritma Naïve Bayes Dengan Metode Syntetic Minority Oversampling Technique (Smote) Pada Data Numerik," *Infotek J. Inform. dan Teknol.*, vol. 8, no. 1, pp. 217–227, 2025, doi: 10.29408/jit.v8i1.28340.

[11] S. M. Natzir, "Perbandingan Kinerja Model Pembelajaran Mesin Dalam Prediksi Banjir Menggunakan Knn, Naive Bayes, Dan Random Forest," *HOAQ (High Educ. Organ. Arch. Qual. J. Teknol. Inf.*, vol. 14, no. 2, pp. 59–64, 2023, doi: 10.52972/hoaq.vol14no2.p59-64.

[12] B. Bahar, "Model Pengujian Akurasi Berbasis Empiris Pada Algoritma Apriori," *Jutisi J. Ilm. Tek. Inform. dan Sist. Inf.*, vol. 8, no. 2, pp. 45–56, 2019.

[13] A. B. Raharjo and M. Quafafou, "Penggabungan Keputusan Pada Klasifikasi Multi-Label," *JUTI J. Ilm. Teknol. Inf.*, vol. 13, pp. 12–23, 2015, doi: 10.12962/j24068535.v13i1.a384.

[14] A. M. Wahid, Turino, K. A. Nugroho, T. Safitri, Darmono, and F. S. Utomo, "Optimasi Logistic Regression dan Random Forest untuk Deteksi Berita Hoax Berbasis TF-IDF," *J. Pendidik. dan Teknol. Indones.*, vol. 4, no. 8, pp. 381–392, 2024, [Online]. Available: https://doi.org/10.52436/1.jpti.602

[15] F. S. Alfiani and U. L. Yuhana, "Implementasi Metode Klasifikasi Multilabel untuk Kategorisasi Materi Pembelajaran Secara Otomatis," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 4, pp. 1750–1758, 2021, doi: 10.35957/jatisi.v8i4.1210.