5th Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)
7 September 2024 – Jakarta, Indonesia
Volume 3, Nomor 2, September 2024 - ISSN 2962-8628 (*online*)

IMPLEMENTASI *NAÏVE BAYES* UNTUK KLASIFIKASI PREDIKSI SERTA ANALISIS DATA BANJIR DI WILAYAH JAKARTA PUSAT

Septian Aji Saputra[1], Hari Soetanto[2*]

[1,2] Teknik Informatika, Teknologi Informasi, Universitas Budi Luhur, Jakarta Selatan, Indonesia

Email: [1]2011501620@student.budiluhur.ac.id, [2*]hari.soetanto@budiluhur.ac.id
(* : corresponding author)

Abstrak-Prediksi adalah proses dalam *data mining* yang melibatkan analisis data untuk meramalkan nilai atau hasil yang belum diketahui berdasarkan pola dan data historis dalam kumpulan data. Metode ini sangat berguna di berbagai sektor, termasuk lingkungan, untuk memperkirakan bencana alam seperti banjir dengan lebih akurat. Jakarta, yang telah mengalami banjir berulang kali selama bertahun-tahun, menghadapi dampak signifikan terhadap kehidupan masyarakat dan memerlukan perhatian khusus dalam upaya mitigasi. Salah satu tantangan utama yang diidentifikasi dalam penelitian ini adalah kesulitan dalam memahami dan mengelola prediksi banjir di Jakarta. Membangun sistem peringatan dini yang efektif berdasarkan hasil prediksi dapat memberikan informasi penting kepada masyarakat dan pemerintah, sehingga memungkinkan tindakan pencegahan dan mitigasi yang tepat waktu. Penelitian ini menggunakan algoritma *naïve bayes* untuk melakukan prediksi banjir. Proses penelitian dimulai dengan pengumpulan 1.793 data, yang setelah tahap preprocessing berkurang menjadi 657 data. Setelah pelabelan, 62 data teridentifikasi sebagai kejadian banjir dan 595 data sebagai tidak banjir. Data kemudian dibagi menjadi dua set: 525 data digunakan untuk pelatihan model, sedangkan 132 data digunakan untuk pengujian. Hasil evaluasi menunjukkan bahwa model *naïve bayes* mencapai akurasi sebesar 97,73%, dengan *precision* 80,00%. *Recall* mencapai 100,00%, dan *f1-score* mencapai 88,98%, yang mengindikasikan bahwa model ini sangat efektif dalam mengklasifikasikan kejadian banjir, memberikan kontribusi penting untuk pengembangan sistem peringatan dini yang lebih baik.

Kata kunci : prediksi, *data mining*, *naïve bayes*, banjir.

IMPLEMENTATION OF NAÏVE BAYES FOR PREDICTION CLASSIFICATION AND ANALYSIS OF FLOOD DATA IN CENTRAL JAKARTA AREA

Abstract-*Prediction is a process in data mining that involves analysing data to forecast unknown values or outcomes based on patterns and historical data in a data set. This method is very useful in various sectors, including the environment, to forecast natural disasters such as floods more accurately. Jakarta, which has experienced repeated flooding over the years, faces significant impacts on people's lives and requires special attention in mitigation efforts. One of the main challenges identified in this research is the difficulty in understanding and managing flood predictions in Jakarta. Building an effective early warning system based on prediction results can provide critical information to the community and government, enabling timely prevention and mitigation actions. This research uses the naïve bayes algorithm to perform flood prediction. The research process began with the collection of 1,793 data, which after the preprocessing stage was reduced to 657 data. After labelling, 62 data were identified as flood events and 595 data as non-flood events. The data was then divided into two sets: 525 data were used for model training, while 132 data were used for testing. The evaluation results showed that the naïve bayes model achieved an accuracy of 97.73%, with a precision of 80.00%. Recall reached 100.00%, and F1-score reached 88.98%, indicating that the model is very effective in classifying flood events, making an important contribution to the development of a better early warning system.***.***

Keywords*: prediction, data mining, naïve bayes, flood.*

1. PENDAHULUAN

Prediksi merupakan suatu proses *data mining* yang menggunakan teknik analisis data untuk memprediksi nilai atau hasil yang belum diketahui berdasarkan pola dan data historis yang ditemukan dalam kumpulan data. Proses ini melibatkan pengolahan dan analisis data secara menyeluruh untuk menemukan tren dan hubungan yang tidak terlihat secara kasat mata. Prediksi dalam *data mining* dapat diterapkan di berbagai bidang, termasuk sektor lingkungan, di mana teknologi ini sangat berguna untuk memprediksi bencana alam seperti banjir dengan lebih akurat. Oleh karena itu, prediksi yang dibuat tidak hanya membantu mencegah dan mengurangi bencana, tetapi juga membantu dalam perencanaan dan pengelolaan sumber daya secara lebih efisien.

Berdasarkan kondisi *geografi*s, DKI Jakarta terletak pada 6° 12” Lintang Selatan dan 106° 48” Bujur Timur. Jakarta adalah dataran rendah dengan ketinggian kurang lebih 7 meter di atas permukaan air laut. 13 sungai

melintasi Jakarta dan bermuara di bagian utara. Sungai ini berperan sebagai salah satu penyebab utama banjir Jakarta. Banjir selalu terjadi di Jakarta selama masa kolonial Belanda dan setelah kemerdekaan. Beberapa peristiwa banjir yang signifikan termasuk banjir tahun 1960 selama masa Presiden Soekarno, banjir besar tahun 1976-1979 yang merusak 1.100 hektare pemukiman, banjir tahun 1985 di Jakarta Barat dan Selatan karena Sungai Pesanggrahan meluap, dan banjir tahun 2007 setelah Orde Baru, ketika 60% wilayah kota kerendam dan 80 orang tewas. Selain itu, banjir-banjir yang terjadi pada tahun 2013, 2015, 2018, 2020, dan 2021 menunjukkan bahwa banjir di Jakarta terus-menerus dan rumit, memengaruhi kehidupan masyarakat, dan membutuhkan perawatan yang serius [1]. Dengan data historis banjir ini, menjadi semakin penting untuk menerapkan prediksi menggunakan *data mining* terhadap banjir di wilayah Jakarta Pusat.

Penelitian ini menggunakan salah satu penelitian sebelumnya sebagai referensi. Studi sebelumnya telah berhasil memprediksi data banjir dengan metode *decision tree, naïve bayes* dan *random fores*. Penelitian ini memberikan dasar yang kuat untuk pengembangan metode prediksi yang lebih akurat untuk menangani masalah banjir. Sebagai contoh penelitian yang dilakukan oleh Muhammad Bagas Arya Darmawan, Favian Dewanta, dan Sri Astuti melakukan penelitian dengan menggunakan *decision tree*, *random forest*, dan *naïve bayes*. Hasilnya menunjukkan bahwa *decision tree* memiliki akurasi tertinggi pada instruksi (99.58%) dan akurasi tertinggi pada pengujian (99.67%). *Random forest* mendominasi pengujian dengan akurasi 99,31% dan akurasi 99,67%. Pada pengujian kedua, dengan rasio 7:3, *decision tree* dan *random forest* mencatat akurasi yang hampir sama (98.87%) dan akurasi pengujian yang hampir sama (98.87%), tetapi *random forest* unggul pada akurasi pengujian dengan 99%. Pada pengujian ketiga, dengan rasio 6:4, Waktu komputasi tercepat adalah rata-rata 0.2 detik oleh *naïve bayes*, diikuti oleh *decision tree* (0.28 detik) dan *random forest* (2-3 detik) [2]. Hasil penelitian tersebut menjadi pedoman dalam pengembangan penelitian berikutnya untuk mencapai efektivitas yang lebih baik.

Penelitian ini berbeda dari studi sebelumnya dengan hanya berfokus khusus pada metode *naïve bayes* untuk prediksi banjir, sedangkan penelitian sebelumnya menggunakan berbagai metode seperti *decision tree*, *random forest*, dan *naïve bayes* secara bersamaan. Meski studi sebelumnya menunjukkan bahwa *decision tree* dan *random forest* memiliki akurasi yang sangat tinggi, penelitian ini mengusulkan pendekatan yang lebih mendalam terhadap metode *naïve bayes* dan menggunakan dataset yang berbeda. Meskipun *naïve bayes* tergolong metode sederhana, penelitian ini bertujuan untuk menguji keandalannya dalam skala yang lebih luas dan memberikan kontribusi baru dalam prediksi banjir dengan teknik *data mining*, menawarkan pendekatan yang lebih terfokus dan inovatif dibandingkan penelitian sebelumnya.

Oleh karena itu, tujuan utama penelitian ini adalah mencapai hasil optimal dalam penerapan metode *naïve bayes* untuk prediksi banjir. Penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam memahami dan menangani masalah prediksi banjir secara lebih efisien.

2. METODE PENELITIAN

Dalam penerapannya, *data mining* melibatkan sejumlah langkah penting yang mempengaruhi hasil akhirnya. Dalam penelitian ini, serangkaian tahapan *data mining* dilakukan untuk mendapatkan hasil yang optimal.

2.1 Pengumpulan Data

Sebelum memulai tahap *data mining*, pemilihan data adalah hal terpenting dalam mengumpukan data operasional yang harus dilakukan. Data yang dipilih ini disimpan dalam *file* yang berbeda dari *database* operasional dan digunakan dalam proses *data mining* [3].

2.2 *Pre-processing*

Pembersihan data atau *Pre-processing* adalah tindakan dasar yang sama seperti penghapusan sebelumnya. Sebelum *data mining* dapat dilakukan, proses pembersihan data adalah fokus setiap penelitian. Proses pembersihan terdiri dari menghilangkan data *null* serta memfilter data yang akan digunakan dalam penelitian [4].

2.3 *Labelling*

Dalam proses pelabelan data, diputuskan untuk mengelompokkan data menjadi dua kategori, yaitu positif dan negatif. Kategori positif meliputi hal-hal atau kegiatan serta ucapan yang secara umum terkait dengan aspek positif. Sementara itu, kategori negatif mencakup hal-hal yang berkaitan dengan kegiatan buruk dan merugikan, serta yang menyebabkan penderitaan. Proses pengelompokan ini dilakukan secara manual dan bersifat subjektif [5]

2.4 *Split* Data

Data yang telah melalui tahap *preprocessing* kemudian digunakan untuk klasifikasi sentimen dengan algoritma *Naive bayes*. Hasil analisis ini akan diklasifikasikan ke dalam dua kategori positif dan negatif. Tahap ini mencakup ekstraksi fitur dari data ulasan, pembagian data menjadi set data latih dan uji, serta pelaksanaan klasifikasi sentimen itu sendiri [6].

2.5 *Modelling*

Pada tahap *Modelling*, dilakukan penerapan teknik-teknik pengklasifikasian data yang paling akurat guna mencapai tujuan penelitian. Proses ini melibatkan pemilihan algoritma yang tepat, pengujian berbagai metode klasifikasi, dan pengoptimalan model untuk memastikan hasil yang optimal. [7].

2.6 Evaluasi Hasil Pengujian

Naïve bayes merupakan metode klasifikasi yang berdasarkan pada *teorama bayes*. Dinamakan *teorema bayes* karena disesuaikan dengan nama penemunya, yaitu Reverend Thomas Bayes, walaupun sebenarnya ada beberapa penelitian yang mengatakan bahwa *teorema bayes* telah ditemukan oleh orang lain sebelum Reverend Thomas Bayes [8]. *Gaussian naïve bayes* merupakan salah satu varian dari algoritma *naïve bayes* yang dirancang untuk menangani data kontinu. Dalam metode ini, diasumsikan bahwa nilai-nilai fitur mengikuti distribusi *Gaussian* (normal). Berikut adalah langkah-langkah dan rumus dasar yang digunakan dalam *Gaussian naïve bayes.*

Probabilitas prior (P(C))
$$P(C_k) = \frac{\text{Jumlah sampel dalam } class\ C_k}{\text{Total jumlah sampel}} \quad (1)$$

Rata-rata *mean*
$$\mu X_i \mathrm{I} C_k = \frac{1}{N_k} \sum_{J=1}^{N_k} x i_j \quad (2)$$

Standar Deviasi (σ)
$$\sigma_{Xi} \mathrm{I} C_k = \sqrt{\frac{1}{N_k} \sum_{j=1}^{N_k} (X_{ij} - \mu_{Xi} \mathrm{I} c_k)^2} \quad (3)$$

Probabilitas *likelihood* (P(X|C))
$$P(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu_{ik})^2}{2\sigma^2}\right) \quad (4)$$

Probabilitas posterior (P(C|X))
$$P(C_k \mathrm{I} X) \infty P(C_k) \prod_{i=1}^{n} P\ (x_i \mathrm{I}\ C_k) \quad (5)$$

Prediksi
$$\hat{C} = \arg\max P\ (C_k \mathrm{I} X) \quad (6)$$

Untuk melakukan evaluasi terhadap model klasifikasi berdasarkan perhitungan objek testing mana yang diprediksi benar dan tidak benar. *Confusion matrix* berisi informasi tentang aktual (*actual*) dan prediksi (*predicted*) pada sistem klasifikasi. Kinerja sistem seperti ini biasanya dievaluasi dengan menggunakan data pada matriks. Perhitungan dimasukan ke dalam tabel *Confusion matrix* [9]. Selain itu, *matrix* ini menilai kinerja klasifikasi berdasarkan apakah objek diklasifikasikan dengan benar atau salah. Setiap bagian menunjukkan jumlah prediksi yang benar dan salah dalam masing-masing kategori. Ini memudahkan kita untuk melakukan perbaikan yang lebih spesifik pada model untuk meningkatkan akurasi dan efektivitas klasifikasi [10].

Tabel 1. *Confusion matrix*

		Predicted Class	
		Banjir	Tidak Banjir
Actual Class	Banjir	(*True positive* = TP)	(*False positive* = FP)
	Tidak Banjir	(*False negative*) = FN)	(*True negative*= TN)

Keterangan:

TP (*True positive*) : Jumlah *class* yang sebenarnya banjir dan diprediksi sebagai banjir oleh model.

FN(*False negative*) : Jumlah *class* yang sebenarnya banjir tetapi diprediksi sebagai tidak banjir oleh model.

FP (*False positive*) : Jumlah *class* yang sebenarnya tidak banjir tetapi diprediksi sebagai banjir oleh model.
TN (*True negative*) : Jumlah *class* yang sebenarnya tidak banjir dan diprediksi sebagai tidak banjir oleh model.

Seperti yang sudah dijelaskan, pengukuran tingkat akurasi, *precision*, *Recall* dan *F1-score* dapat diketahui melalui *Confusion matrix* dengan penjelasan rumus sebagai berikut:

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

$$\textit{Precision} = \frac{TP}{TP+FP} \quad (8)$$

$$\textit{Recall} = \frac{TP}{TP+FN} \quad (9)$$

$$\textit{F1-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (10)$$

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Data untuk penelitian kali ini diperoleh dari situs resmi BMKG, url : https://dataonline.bmkg.go.id/data_iklim, yang meliputi atribut temperatur rata-rata, yang mengukur tingkat panas atau dinginnya suatu lingkungan atau objek, sering kali dinyatakan dalam derajat Celsius (°C). Selain itu, ada curah hujan, yang mengukur jumlah air hujan yang jatuh di suatu wilayah dalam periode waktu tertentu, umumnya diukur dalam milimeter (mm). Di samping itu, kecepatan angin juga merupakan atribut yang mencerminkan seberapa cepat udara bergerak di suatu area, biasanya diukur dalam kilometer per jam (km/jam) atau meter per detik (m/s). Di samping atribut-atribut tersebut, terdapat juga atribut yang berasal dari situs resmi portal data Dinas Sumber Daya Air Provinsi DKI Jakarta., url : https://portaldatadsda.jakarta.go.id/. Meliputi atribut seperti kedalaman kali ciliwung, yang diukur dari permukaan air hingga dasar sungai dalam satuan sentimeter (cm), serta lebar kali ciliwung, yang mengukur jarak horizontal dari satu sisi sungai ke sisi lainnya juga dalam sentimeter (cm).. Untuk total data yang dikumpulkan berjumlah 1.793 *record* data. Dibawah ini merupakan sampel data banjir dengan jumlah 10 *record*.

Tabel 2. Sampel data banjir

Temperatur rata-rata	Curah hujan	Kecepatan angin	Lebar maksimum kali ciliwung	Kedalaman maksimum kali ciliwung
29.4	5.6	2	50	12
28	0	2	50	12
27.5		2	50	12
26.8	145.3	1	50	12
27.8	12.9	1	50	12
28.3	0	1	50	12
28.4		2	50	12
27.8	0	1	50	12
27.7	70.4	1	50	12
29.1	50.6	2	50	12

3.2 *Pre-processing*

pre-processing, yang memiliki tujuan untuk menyediakan data mentah dalam format yang bersih dan siap digunakan untuk pemodelan atau analisis lebih lanjut. Dalam penelitian kali ini, proses ini melibatkan beberapa langkah, dimulai dengan pembersihan data (*cleaning* data) dan penyaringan data (*filtering* data).

a. *Cleaning* data

Pada tahap ini, nilai-nilai tertentu seperti 8888, yang sebelumnya digunakan untuk menunjukkan data yang tidak terhitung, diubah menjadi nilai *null*. Setelah nilai-nilai tersebut diubah menjadi *null*, data kemudian dibersihkan dari semua nilai *null* untuk memastikan keakuratan dataset.

Tabel 3. Sampel proses *cleaning*

Temperatur rata-rata	Curah hujan	Kecepatan angin	Lebar maksimum kali ciliwung	Kedalaman maksimum kali ciliwung
29.4	5.6	2	50	12
28	0	2	50	12
26.8	145.3	1	50	12
27.8	12.9	1	50	12
28.3	0	1	50	12
27.8	0	1	50	12
27.7	70.4	1	50	12
29.1	50.6	2	50	12

b. *Filtering* data

Dilakukan proses penyaringan untuk memilih data curah hujan yang > 0 mm/hari. Langkah ini bertujuan untuk memastikan bahwa hanya data yang relevan, yaitu data dengan curah hujan > 0 mm/hari.

Tabel 4. Sampel proses *filtering*

Temperatur rata-rata	Curah hujan	Kecepatan angin	Lebar maksimum kali ciliwung	Kedalaman maksimum kali ciliwung
29.4	5.6	2	50	12
26.8	145.3	1	50	12
27.8	12.9	1	50	12
27.7	70.4	1	50	12
29.1	50.6	2	50	12

3.3 *Labelling* data

Label ini memiliki data parameter curah hujan sebagai acuan. Setelah itu, tentukan label yang relevan, seperti "Banjir" untuk curah hujan yang >=40 mm/hari dan "Tidak Banjir" untuk <40 mm/hari dari parameter curah hujan. Dasar pelabelan banjir ini didasarkan pada diskusi mendalam dengan pakar banjir serta melakukan analisis data banjir di wilayah Jakarta Pusat. Proses ini mempertimbangkan curah hujan yang signifikan, yaitu >=40mm/hari, sebagai indikator utama dalam mengklasifikasikan potensi banjir. Namun, kelemahan dari pendekatan ini adalah fokusnya yang terbatas pada analisis data banjir dan curah hujan, tanpa memperhitungkan data tambahan dari BMKG atau instansi lain yang relevan dengan sektor banjir.

Tabel 5. Sampel proses *labelling*

Temperatur rata-rata	Curah hujan	Kecepatan angin	Lebar maksimum kali ciliwung	Kedalaman maksimum kali ciliwung	Kategori
29.4	5.6	2	50	12	Tidak banjir
26.8	145.3	1	50	12	Banjir
27.8	12.9	1	50	12	Tidak banjir
27.7	70.4	1	50	12	Banjir
29.1	50.6	2	50	12	Banjir

3.4 *Split* data

Setelah *labelling* pada titik ini, data akan dibagi menjadi dua bagian yakni data latih dan data uji. Delapan puluh persen dari bagian ini dialokasikan untuk data latih dan dua puluh persen lagi dialokasikan untuk data uji. Tujuannya adalah untuk mengoptimalkan hasil analisis dengan memastikan model yang dibangun dapat digeneralisasi dengan baik.

a. Data Latih

Setelah proses *split* data dilakukan, langkah-langkah sampel menghasilkan data latih yang digunakan untuk melatih model. Data latih, yang merupakan sebagian besar dataset awal, digunakan untuk membangun dan mengoptimalkan model. Berikut adalah *detail* data latih yang dihasilkan.

Tabel 6. Sampel data latih

Temperatur rata-rata	Curah hujan	Kecepatan angin	Lebar maksimum kali ciliwung	Kedalaman maksimum kali ciliwung	Kategori
29.4	5.6	2	50	12	Tidak banjir
27.7	70.4	1	50	12	Banjir

29.1	50.6	2	50	12	Banjir

Data diatas merupakan sampel data latih yang telah melalui proses split data.

b. Data Uji

Selain data latih, data uji juga merupakan data yang telah melalui proses *split* data untuk mendapatkan data untuk di lakukan pengujian.

Tabel 7. Sampel data uji

Temperatur rata-rata	Curah hujan	Kecepatan angin	Lebar maksimum kali ciliwung	Kedalaman maksimum kali ciliwung	Kategori
26.8	145.3	1	50	12	Banjir
27.8	12.9	1	50	12	Tidak banjir

Data diatas merupakan sampel data uji yang telah melalui proses split data.

3.5 *Modelling*

Setelah proses pembagian data selesai, langkah berikutnya adalah memodelkan data latih. Pada tahap ini, algoritma pemodelan digunakan untuk membuat model yang dapat memprediksi atau mengklasifikasikan data sesuai dengan pola yang ditemukan dalam data latih.

3.6 Evaluasi Pengujian

Setelah pemodelan selesai, langkah selanjutnya adalah evaluasi hasil sampel data uji. Proses ini menghasilkan nilai akurasi, *precision*, *Recall*, dan *F1-score*. Selain itu ada langkah tambahan untuk menghitung nilai probabilitas saat menilai hasil.

Tabel 8. Sampel data untuk evaluasi pengujian

Temperatur rata-rata	Curah hujan	Kecepatan angin	Lebar maksimum kali ciliwung	Kedalaman maksimum kali ciliwung	Kategori
27.7	70.4	1	50	12	Banjir
27.8	12.9	1	50	12	Tidak Banjir
29.1	50.6	2	50	12	Banjir
29.4	5.6	2	50	12	Tidak Banjir

Setelah memilih sampel data untuk pengujian, langkah berikutnya adalah menyiapkan *confusion matrix*. Ini dilakukan untuk mengevaluasi bagaimana model dapat memprediksi data dengan membandingkan hasil prediksi dengan nilai sebenarnya dari sampel yang diuji.

Tabel 9. Sampel *confusion matrix* pada evaluasi pengujian.

		Predicted Class	
		Banjir	Tidak Banjir
Actual Class	Banjir	2	0
	Tidak Banjir	0	2

Setelah memahami rumus *confusion matrix*, berikut adalah langkah-langkah yang dilakukan untuk melaklukan evaluasi hasil. Langkah pertama kita adalah dimana kita perlu menghitung nilai akurasi, *precision*, *Recall*, dan *F1-score*. Proses ini melibatkan penerapan rumus yang telah dijelaskan pada bab sebelumnya.

Tabel 10. Sampel hasil perhitungan *confusion matrix.*

Akurasi	*Precision*	*Recall*	*F1-score*
100%	100%	100%	100%

Setelah mendapatkan nilai akurasi, *precision*, *Recall*, dan *F1-score*, langkah berikutnya adalah menghitung nilai probabilitas menggunakan algoritma *gaussian naïve bayes.*

a. Menghitung probabilitas prior (P(C)).

Langkah pertama dalam menghitung probabilitas prior (P(C)) menggunakan rumus persamaan yang ditemukan pada bab sebelumnya. dimana setiap kelas harus diharus dipisah untuk dilakukan perhitungan. Untuk menghitung probabilitas prior (P(C)), kita dapat menggunakan jumlah kemunculan setiap kategori dibagi dengan total jumlah data.

Tabel 11. Sampel data banjir evaluasi pengujian.

Temperatur rata-rata	Curah hujan	Kecepatan angin	Lebar maksimum kali ciliwung	Kedalaman maksimum kali ciliwung	Kategori
27.7	70.4	1	50	12	Banjir
29.1	50.6	2	50	12	Banjir

Kedua data tersebut merupakan sampel data yang masuk ke dalam *class* banjir.

Tabel 12. Sampel data tidak banjir evaluasi pengujian.

Temperatur rata-rata	Curah hujan	Kecepatan angin	Lebar maksimum kali ciliwung	Kedalaman maksimum kali ciliwung	Kategori
27.8	12.9	1	50	12	Tidak Banjir
29.4	5.6	2	50	12	Tidak Banjir

Kedua data tersebut merupakan sampel data yang masuk ke dalam *class* tidak banjir.

Tabel 13. Sampel hasil probabilitas prior

Probabilitas prior banjir	Probabilitas prior tidak banjir
0.5	0.5

Nilai diatas merupakan nilai probabilitas prior dari masing-masing *class.*

b. Menghitung *mean* dan standar deviasi.

Setelah mendapatkan nilai probabilitas prior (P(C)), langkah selanjutnya menghitung nilai rata-rata *mean* serta standar deviasi dengan persamaaan rumus pada bab sebelumnya.

Tabel 14. Sampel hasil rata-rata *mean* dan standar deviai data banjir.

Atribut	*Mean*	Standar Deviasi
Temperatur rata-rata	28.4	0.70
Curah hujan	60.5	9.90
Kecepatan angin	1.5	0.50
Kedalaman kali ciliwung	50	0.01
Lebar kali ciliwung	12	0.01

Hasil di atas merupakan nilai rata-rata *mean* dan standar deviasi dari kedua atribut untuk *class* banjir.

Tabel 15. Sampel hasil rata-rata *mean* dan standar deviasi data tidak banjir.

Atribut	*Mean*	Standar Deviasi
Temperatur rata-rata	28.6	0.80
Curah hujan	9.2	3.63
Kecepatan angin	1.5	0.50
Kedalaman kali ciliwung	50	0.01
Lebar kali ciliwung	12	0.01

Hasil di atas merupakan nilai rata-rata *mean* dan standar deviasi dari kedua atribut untuk *class* tidak banjir.

c. Menghitung probabilitas *likelihood.*

Setelah mendapatkan nilai rata-rata (*mean*) dan standar deviasi dari data yang terkait dengan banjir dan tidak banjir, langkah selanjutnya adalah menghitung probabilitas *likelihood.*

Tabel 16. Sampel hasil *likelihood* dari data *class* banjir.

Sampel data banjir	Total *Likelihood* banjir
Sampel data banjir 1	6.500824581
Sampel data banjir 2	6.500824581

Berdasarkan hasil sampel data *class* banjir *likelihood* tersebut, hasil ini diperoleh saat menerapkan rumus persamaan pada bab sebelumnya.

Tabel 17. Sampel hasil *likelihood* dari data *class* tidak banjir.

Sampel data tidak banjir	Total *Likelihood* banjir
Sampel data tidak banjir 1	15.42785922
Sampel data tidak banjir 2	15.42785922

Berdasarkan hasil sampel data *class* tidak banjir *likelihood* tersebut, hasil ini diperoleh saat menerapkan rumus persamaan pada bab sebelumnya.

d. Menghitung probabilitas posterior.

Setelah mendapatkan nilai probabilitas *likelihood* dari kedua *class*, langkah selanjutnya adalah menghitung nilai probabilitas posterior menggunakan persamaan yang dijelaskan pada bab sebelumnya. Sebelum mencapai hasil akhir probabilitas posterior, langkah-langkahnya meliputi menghitung nilai log probabilitas prior, menghitung rata-rata dari total *likelihood* setiap kelas, menghitung log dari rata-rata *likelihood* tersebut, dan akhirnya, mencapai perhitungan probabilitas posterior.

Tabel 18. Sampel hasil nilai log probabilitas prior.

	Banjir	Tidak Banjir
Probabilitas prior	0.5	0.5
Log probabilitas prior	- 0.30103	- 0.30103

Hasil di atas merupakah sampel nilai log pada probabilitas prior dari masing-masing *class*.

Tabel 19. Sampel hasil nilai log probabilitas *likelihood*.

	Banjir	Tidak Banjir
Probabilitas *likelihood*	6.500824581	15.42785922
Log probabilitas *likelihood*	0.812968447	1.188305667

Hasil di atas merupakah sampel nilai log pada probabilitas *likelihood* dari masing-masing *class*.

Tabel 20. Sampel hasil perhitungan probabilitas posterior.

	Banjir	Tidak Banjir
Log probabilitas prior	- 0.30103	- 0.30103
Log probabilitas *likelihood*	0.812968447	1.188305667
Hasil probabilitas posterior	0.511938451	0.887275672

Hasil akhir di atas adalah penjumlahan log probabilitas prior dengan log probabilitas *likelihood*, yang menunjukkan bahwa hasil tersebut merupakan nilai dari probabilitas posterior.

e. Prediksi

Berdasarkan hasil probabilitas posterior tersebut, diperkirakan bahwa dari data sampel evaluasi pengujian tersebut, probabilitas pada *class* tidak banjir lebih tinggi dibandingkan dengan *class* banjir.

3.7 Pengujian

Pengujian adalah tahap yang sangat penting dalam proses pengembangan dan penyempurnaan aplikasi. Melalui pengujian, kita dapat memastikan aplikasi berfungsi sebagaimana mestinya dan sesuai dengan kebutuhan pengguna. dalam penelitian ini peneliti menggunakan data sebanyak 525 untuk data latih dan 132 untuk data uji. Pengujian yang dilakukan dalam penelitian ini dilakukan menggunakan rancangan yang dibuat pada bab sebelumnya. Untuk mengevaluasi kinerja aplikasi dalam mendeteksi dan memprediksi data banjir, berikut adalah pengujian yang menunjukkan hasil akurasi, *precision*, *Recall*, dan *F1-score* pada tabel *confusion matrix*.

Tabel 21. Pengujian dari keseluruhan data uji pada *confusion matrix*

		Predicted Class	
		Banjir	Tidak Banjir
Actual Class	Banjir	12	0
	Tidak Banjir	3	117

Hasil evaluasi pengujian menunjukkan performa yang sangat baik berdasarkan tabel *confusion matrix* yang diberikan. *True Positives* (TP) sebanyak 12 menunjukkan bahwa pengujian tersebut antara aktual banjir dan prediksi banjir sesuai. *False Negatives* (FN) berjumlah 3, mengindikasikan bahwa ada 3 kasus banjir yang tidak terdeteksi oleh model. *False Positives* (FP) tidak ada dalam *confusion matrix* adalah indikasi yang sangat positif karena antara model dengan aktual sesuai. *True Negatives* (TN) sebanyak 117 menunjukkan bahwa model mampu mengklasifikasikan 117 kasus tidak banjir dengan benar.

Tabel 22. Hasil akhir pengujian pada data uji

Akurasi	*Precision*	*Recall*	*F1-score*
97.73%	80.00%	100.00%	88.89%

Hasil pengujian mengindikasikan bahwa model memiliki kinerja yang sangat memuaskan dengan akurasi sebesar 97,73%. Hasil tersebut diperoleh setelah melalui proses pengumpulan dataset yang luas serta penerapan preprocessing data yang optimal. Selain itu hasil *Precision* sebesar 80,00% mengindikasikan sebagian besar prediksi positif model adalah benar, sedangkan *Recall* 100,00% menunjukkan model berhasil mengidentifikasi semua kasus positif. *F1-score* 88,89% mencerminkan keseimbangan kuat antara *precision* dan *Recall*, mengindikasikan performa model yang konsisten dan andal.. Hasil tersebut telah melalui tahapan diskusikan dengan ahli yang membahas banjir dan data untuk dianalisis khususnya untuk data banjir di Jakarta Pusat.

4. KESIMPULAN

Berdasarkan hasil penelitian tersebut didapatkan kesimpulan bahwa penelitian ini menggunakan metode *naïve bayes* untuk mengembangkan dan menguji klasifikasi serta prediksi data banjir di Jakarta Pusat. Proses penelitian mencakup tahapan menyeluruh mulai dari pengumpulan data, *pre-processing*, *labelling*, *split* data, hingga pembuatan model dan pengujian. Dari 1.793 data awal, setelah melalui tahap pre-processing, jumlahnya berkurang menjadi 657 data, dengan 62 di antaranya diidentifikasi sebagai kejadian banjir dan 595 sebagai tidak banjir. Data ini kemudian dibagi menjadi 525 data latih dan 132 data uji. Hasilnya, model menunjukkan kinerja luar biasa dengan akurasi mencapai 97,73%, *precision* 80,00%, *Recall* 100,00%, dan *F1-score* 88,98%, yang mengindikasikan kemampuan model untuk mengklasifikasikan kejadian banjir dengan sangat baik. Untuk penelitian ini, diperlukan keahlian khusus mengenai banjir dan analisis data banjir di wilayah Jakarta Pusat. Beberapa saran untuk pengembangan lebih lanjut dari penelitian ini meliputi, penambahan atribut ketinggian muka air dan debit air sebagai variabel tambahan, perluasan dataset banjir untuk hasil yang lebih optimal, uji coba dengan metode *data mining* lain untuk perbandingan performa, serta pengumpulan data banjir dari institusi pemerintah sebagai parameter klasifikasi yang lebih akurat

DAFTAR PUSTAKA

[1] N. Y. Puspita, F. Sembiring, and A. R. H. Putra, "Mitigasi Banjir Pada Saat Pandemi Covid-19: Sudah Siapkah Pemerintah DKI Jakarta?," *J. Pendidik. Kewarganegaraan Undiksha*, vol. 10, no. 1, pp. 129–146, 2022, [Online]. Available: https://ejournal.undiksha.ac.id/index.php/JJPP

[2] M. B. Arya Darmawan, F. Dewanta, and S. Astuti, "Analisis Perbandingan Algoritma Decision Tree, Random Forest, dan *Naïve bayes* untuk Prediksi Banjir di Desa Dayeuhkolot," *TELKA - Telekomun. Elektron. Komputasi dan Kontrol*, vol. 9, no. 1, pp. 52–61, 2023, doi: 10.15575/telka.v9n1.52-61.

[3] W. Romadhona, B. Indarmawan Nugroho, and A. Alim Murtopo, "Implementasi *Data mining* Pemilihan Pelanggan Potensial Menggunakan Algoritma K-Means," *J. Minfo Polgan*, vol. 11, no. 2, pp. 100–104, 2022, doi: 10.33395/jmp.v11i2.11797.

[4] W. Kokoh Andriyan, et al, "Penerapan *Data mining* Dengan Menggunakan Metode K-Means Clustering Dalam Pengelompokan Data Nilai Pada SMA YKPP Pendopo Untuk Menentukan Jurusan IPA dan IPS," *J. Jupiter*, vol. 15, no. 1, pp. 452–461, 2023.

[5] D. Oktavia, Y. R. Ramadahan, and Minarto, "Analisis Sentimen Terhadap Penerapan Sistem E-Tilang Pada Media Sosial Twitter Menggunakan Algoritma Support Vector Machine (SVM)," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 1, pp. 407–417, 2023, doi: 10.30865/klik.v4i1.1040.

[6] Ernianti Hasibuan and Elmo Allistair Heriyanto, "Analisis Sentimen Pada Ulasan Aplikasi Amazon Shopping Di Google Play Store Menggunakan Naive Bayes Classifier," *J. Tek. dan Sci.*, vol. 1, no. 3, pp. 13–24, 2022, doi: 10.56127/jts.v1i3.434.

[7] D. Pratmanto and F. F. D. Imaniawan, "Analisis Sentimen Terhadap Aplikasi Canva Menggunakan Algoritma Naive Bayes Dan K-Nearest Neighbors," *Comput. Sci.*, vol. 3, no. 2, pp. 110–117, 2023, doi: 10.31294/coscience.v3i2.1917.

[8] R. Amalia, "Penerapan *Data mining* Untuk Memprediksi Hasil Kelulusan Siswa menggunakan Metode *Naïve bayes*," *J. Inform. dan Sist. Inf.*, vol. 6, no. 1, pp. 33–42, 2020.

[9] A. Ridwan, "Penerapan Algoritma *Naïve bayes* Untuk Klasifikasi Penyakit Diabetes Mellitus," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 4, no. 1, pp. 15–21, 2020, doi: 10.47970/siskom-kb.v4i1.169.

[10] E. S. Ilna Nuzla Putri, Rachmadita Andeswari, "Analisis Dan Deteksi Fraud Pada Data Panggilan Menggunakan Algoritma *Naïve Bayes* Pada PT XYZ," vol. 14, no. 2, pp. 1–15, 2020.