Computer Science (CO-SCIENCE)
Vol. 5 No. 2 Juli 2025
Akreditasi Sinta 4 Nomor SK : 230/E/KPT/2022

Penerapan Integrasi Algoritma *K-Means* Dan *Naïve Bayes* Untuk Klasifikasi Wilayah Rawan Banjir Di Jakarta

Irfan Maulana Sinatrya[1], Achmad Baroqah Pohan[2*], Yunita[3], Hilda Amalia[4], Ade Fitria Lestari[5]

[1,2,3,5]Program Studi Sistem Informasi, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika
Jl. Kramat Raya No.98, RT.2/RW.9, Kwitang, Kec. Senen, Kota Jakarta Pusat, Daerah Khusus Ibukota Jakarta 10450, Indonesia

[4]Program Studi Teknologi Informasi, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika
Jl. Kramat Raya No.98, RT.2/RW.9, Kwitang, Kec. Senen, Kota Jakarta Pusat, Daerah Khusus Ibukota Jakarta 10450, Indonesia

e-mail: [1]Irfmlnsntrya@gmail.com, [2]achmad.abq@bsi.ac.id , [3]yunita.ynt@bsi.ac.id , [4]hilda.ham@bsi.ac.id , [5] ade.afr@bsi.ac.id

(*) Corresponding Author

Artikel Info : Diterima : 22-10-2024 | Direvisi : 16-05-2025 | Disetujui : 20-05-2025

Abstrak - Jakarta, sebagai kota metropolitan di Indonesia, sering mengalami banjir yang disebabkan oleh curah hujan tinggi, sistem drainase yang buruk, dan urbanisasi yang cepat. Penelitian ini bertujuan untuk mengklasifikasikan wilayah rawan banjir di Jakarta menggunakan kombinasi algoritma *K-Means Clustering* dan *Naïve Bayes Classifier*. Tahapan penelitian dimulai dari pengumpulan data dari *website* Satu Data Jakarta, mencakup atribut seperti wilayah, kecamatan, kelurahan, jumlah rata-rata ketinggian air, jumlah RW terdampak, jumlah KK terdampak, jumlah jiwa terdampak, dan jumlah kejadian banjir. Data yang terkumpul kemudian diproses melalui tahap pembersihan dan normalisasi sebelum dianalisis menggunakan algoritma *K-Means* untuk mengelompokkan wilayah berdasarkan karakteristik banjirnya. Selanjutnya, algoritma *Naïve Bayes* digunakan untuk membangun model klasifikasi yang memprediksi wilayah rawan banjir. Hasil penelitian menunjukkan bahwa kombinasi kedua algoritma ini menghasilkan rata-rata akurasi yang lebih tinggi dibandingkan dengan penggunaan *Naïve Bayes* konvensional, memiliki akurasi mencapai 98.18%% pada rasio split data data pelatihan dan pengujian 70:30, 80;20 dan 90:10. Temuan ini memberikan wawasan berharga untuk mitigasi risiko banjir di Jakarta, membantu pemerintah dalam mengambil langkah preventif yang lebih efektif.

Kata Kunci : Banjir, *Data Mining, Naïve Bayes, K-Means*.

Abstracts *- Jakarta, as a metropolitan city in Indonesia, often experiences flooding caused by high rainfall, poor drainage systems, and rapid urbanization. This research aims to classify flood-prone areas in Jakarta using a combination of K-Means Clustering and Naïve Bayes Classifier algorithms. The research phase begins with data collection from the Satu Data Jakarta website, including attributes such as region, sub-district, village, average water level, number of affected RWs, number of affected families, number of affected people, and number of flood events. The collected data is then processed through cleaning and normalization stages before being analyzed using the K-Means algorithm to group areas based on their flooding characteristics. Furthermore, the Naïve Bayes algorithm was used to build a classification model that predicts flood-prone areas. The results showed that the combination of these two algorithms resulted in higher average accuracy compared to the use of conventional Naïve Bayes, having an accuracy of 98.18%% at training and testing data split ratios of 70:30, 80;20 and 90:10. The findings provide valuable insights for flood risk mitigation in Jakarta, assisting the government in taking more effective preventive measures.*

Keywords : Flood, Data Mining, Naïve Bayes, K-Means.

PENDAHULUAN

Negara Kesatuan Republik Indonesia sebagai negara kepulauan secara geografis terletak di garis khatulistiwa, dan berada pada pertemuan tiga lempeng tektonik yaitu lempeng Eurasia, lempeng Indo-Australia

dan lempeng pasifik. Kondisi ini menjadikan Indonesia sebagai wilayah teritorial yang sangat rawan terhadap bencana alam salah satunya adalah banjir. Pemerintah Indonesia telah menetapkan dalam UU nomor 24 tahun 2007 mengenai BPBD (Badan Penanggulangan Bencana Daerah) (Angreini & Supratman, 2021).Banjir didefinisikan oleh Badan Nasional Penanggulangan Bencana (BNPB) sebagai peningkatan volume air yang menggenangi daratan. Banjir merupakan salah satu bencana alam yang paling sering terjadi, dengan proporsi mencapai 40% dari seluruh bencana alam lainnya. Diakibatkan berbagai faktor seperti intensitas curah hujan, kemiringan lereng, aliran limpasan sungai, serta faktor manusia seperti kurangnya perhatian terhadap pelestarian lingkungan sekitar. (Anggraini et al., 2021).

Berdasarkan data yang dimiliki Indeks Risiko Bencana Indonesia (IRBI) tahun 2023, provinsi DKI Jakarta memiliki nilai 61.35 (sedang). Dari nilai tersebut wilayah Jakarta juga memiliki ancaman bencana seperti: gempa bumi, banjir, tanah longsor, kekeringan, cuaca ekstrim, gelombang ekstrem/abrasi. Faktor-faktor seperti kota Jakarta terletak pada topografi yang rendah, perubahan iklim dan pembangunan yang sangat pesat, ditambah dengan kepadatan penduduk yang tinggi, membuat kota Jakarta memiliki potensi bencana banjir yang tinggi (Badan Nasional Penanggulangan Bencana, 2023). oleh sebab itu, untuk mendukung upaya mitigasi resiko kerawanan banjir di Jakarta diperlukan peran teknologi dan pendekatan yg sistematis untuk memberikan masukan kepada pemerintah agar dapat mengambil kebijakan untuk menanggulangi banjir khususnya pada wilayah Jakarta yang sering terdampak banjir.

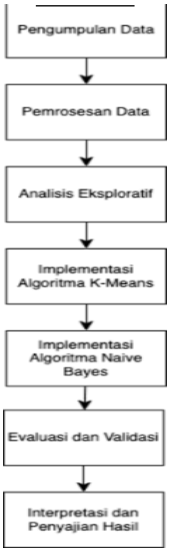
Proses data mining memiliki beberapa metode analisis data, seperti Klustering, Asosiasi, Klasifikasi, Regresi dan Peramalan (Bui & Bahtiar, 2024). Salah satu metode data mining yang sering dipakai untuk mengelompokkan (klustering) data adalah algoritma K-Means. Meminimalkan fungsi objektif yang ditetapkan dalam proses pengelompokan (klustering) adalah tujuan dari pengelompokan. Secara umum, ini berarti meminimalkan variasi dalam suatu kelompok dan memaksimalkan variasi antar kelompok (Khomsiyah et al., 2021). Klasifikasi K-means memainkan peran yang penting dalam menganalisis dan mengeksploratif data dan dapat menentukan model, fungsi atau membedakan konsep dan kelas data dan sangat tepat untuk memprediksi kelas dari suatu objek yang kelasnya tidak diketahui (Effendi et al., 2024).

Naïve Bayes termasuk salah satu proses data mining yang umum digunakan dalam klasifikasi data. Metode *Naïve Bayes* merupakan salah satu algoritma klasifikasi yang efektif, sederhana, efisien, dan dapat diandalkan dalam mengatasi masalah data seperti atribut yang kurang atau hilang serta tidak memerlukan jumlah data yang terlalu besar (Fatonah et al., 2021). Klasifikasi dapat dideskripsikan sebagai metode untuk membuktikan sebuah objek data sebagai salah satu jenis yang telah dideskripsikan sebelumnya (Alghifari & Juardi, 2021). Untuk mendapatkan model pada dataset yang membedakan label yang bersesuaian, model tersebut kemudian digunakan untuk mengklasifikasikan atribut kelas atau label yang tidak diketahui.

Dalam penelitian ini dilakukan implementasi integrasi algoritma *K-Means Clustering* dan *Naïve bayes Classifier*. Pendekatan metode ini dibentuk dengan menggabungkan teknik klustering dan klasifikasi yang nantinya hasil penggabungan metode tersebut akan dibandingkan dengan *Naïve Bayes* konvensional untuk klasifikasi wilayah rawan banjir di kota Jakarta (Nandang Iriadi et al., 2020). Penelitian sebelumnya yang dilakukan oleh (Martin Saputra, 2025). pada penelitian ini memanfaatkan integrasi algoritma dan *Naïve bayes Classifier* untuk melakukan Klasifikasi banjir. Hasil penelitian menujukkan bahwa teknik data mining, khususnya Naïve Bayes dan K-Means clustering, mampu memprediksi banjir di Jakarta pada tahun 2025 dengan tingkat akurasi yang tinggi, mencapai hingga 97%. Analisis data dari sepuluh tahun terakhir digunakan untuk mengklasifikasikan tingkat air dan mengelompokkan daerah yang terdampak. Hasil prediksi menunjukkan adanya kemungkinan penurunan dampak banjir di masa depan.

METODE PENELITIAN

Pada penelitian ini, penulis akan melakukan pendekatan integrasi metode *K-Means Clustering* dan *Naïve Bayes Classifier* dalam mengklasifikasi wilayah rawan banjir di kota Jakarta. Berikut ini merupakan tahapan metode penelitian:



Sumber: Hasil Penelitian (2024)
Gambar 1. Tahapan metode penelitian

Tahapan pertama dalam penelitian ini adalah mengumpulkan data yang relevan untuk menganalisis wilayah rawan banjir di Jakarta. Data yang dikumpulkan mencakup informasi tentang wilayah, kecamatan, kelurahan, jumlah rata-rata ketinggian air, jumlah rw terdampak, jumlah kk terdampak, jumlah jiwa terdampak, terdampak, jumlah kejadian dan faktor-faktor lain yang dapat mempengaruhi risiko banjir. Setelah data terkumpul, tahap selanjutnya adalah pemrosesan data. Ini mencakup pembersihan data untuk menghilangkan data yang tidak relevan atau tidak valid, transformasi data jika diperlukan, dan pemilihan fitur untuk mempersempit cakupan analisis. Pemrosesan data sangat penting untuk menjamin bahwa data yang digunakan dalam penelitian memiliki kualitas yang baik dan siap untuk dianalisis lebih lanjut. Setelah data diproses, tahap selanjutnya adalah analisis eksploratif. Pada tahap ini, peneliti menganalisis data untuk memahami pola-pola yang mungkin ada dan hubungan antara variabel-variabel yang relevan (Zhang, 2020). Analisis ini membantu dalam mengidentifikasi wilayah-wilayah yang rentan terhadap banjir dan faktor-faktor yang mempengaruhinya. Setelah pemahaman yang lebih dalam tentang data tercapai, penelitian melanjutkan dengan implementasi algoritma *K-means*. Algoritma ini digunakan untuk mengelompokkan wilayah-wilayah berdasarkan karakteristik yang serupa terkait risiko banjir. Penggunaan algoritma ini bertujuan untuk mengidentifikasi pola-pola spasial yang mungkin tidak terlihat secara langsung dari data mentah. Selain *K-means*, penelitian juga melibatkan implementasi algoritma *Naïve Bayes*. Algoritma ini digunakan untuk membangun model klasifikasi yang dapat memprediksi wilayah-wilayah yang rentan terhadap banjir berdasarkan atribut-atribut yang diberikan. *Naïve Bayes* digunakan untuk memahami faktor-faktor yang berkontribusi terhadap risiko banjir dan memperkirakan probabilitas kejadian banjir di wilayah-wilayah tertentu. Tahap berikutnya meliputi evaluasi dan validasi model yang telah dibuat dengan menggunakan algoritma K-means dan Naïve Bayes. Evaluasi bertujuan untuk menilai seberapa baik model tersebut dalam mengklasifikasikan daerah-daerah yang rawan banjir. Beberapa metrik evaluasi yang sering dipakai antara lain akurasi, performance vector, dan weighted mean recall (Ridwan, 2020). Sedangkan validasi dilakukan untuk memastikan bahwa model tersebut mampu bekerja dengan baik dan dapat diterapkan pada data baru yang belum pernah digunakan sebelumnya. Terakhir, hasil dari penelitian ini diinterpretasikan dan disajikan secara tepat kepada pemangku kepentingan yang relevan, seperti pemerintah daerah, LSM, dan masyarakat umum. Interpretasi hasil membantu dalam pemahaman tentang wilayah-wilayah rawan banjir di kota Jakarta dan memberikan wawasan yang berharga untuk pengambilan keputusan terkait mitigasi risiko banjir.

HASIL DAN PEMBAHASAN

1. Pengumpulan Data

Dataset dengan nama Data Kejadian Bencana Banjir Tahun 2024 diperoleh dari *website* https://satudata.jakarta.go.id/open-data/detail?kategori=dataset&page_url=data-kejadian-bencana-banjir&data_no=1 berjumlah 158 data, data tersebut memiliki atribut wilayah, kecamatan, kelurahan, jumlah rata-rata ketinggian air, jumlah rw terdampak, jumlah kk terdampak, jumlah jiwa, terdampak, jumlah kejadian. Penelitian ini menggunakan data dari *website* Satu Data Jakarta yang memuat informasi terkait kejadian banjir di wilayah kota Jakarta. *Dataset* yang digunakan memiliki beberapa atribut penting yang akan dianalisis lebih lanjut menggunakan metode *K-Means Clustering* dan *Naïve Bayes Classifier*. Berikut adalah penjelasan atribut-atribut dalam dataset tersebut:

1. Wilayah: Nama wilayah di kota Jakarta yang terdampak banjir.
2. Kecamatan: Nama kecamatan di kota Jakarta yang terdampak banjir.
3. Kelurahan: Nama kelurahan di kota Jakarta yang terdampak banjir.

4. Jumlah Rata-rata Ketinggian Air (cm): Rata-rata ketinggian air yang menggenangi wilayah dalam satuan centimeter.
5. Jumlah RW Terdampak: Jumlah Rukun Warga (RW) yang terdampak oleh banjir.
6. Jumlah KK Terdampak: Jumlah Kepala Keluarga (KK) yang terdampak oleh banjir.
7. Jumlah Jiwa Terdampak: Jumlah jiwa yang terdampak oleh banjir.
8. Terdampak: Indikator apakah wilayah tersebut terdampak oleh banjir atau tidak.
9. Jumlah Kejadian: Jumlah kejadian banjir yang terjadi di wilayah tersebut.

Berikut adalah tabel yang menunjukkan sebagian data yang digunakan dalam penelitian ini:

wilayah	kecamatan	kelurahan	jumla	jumla	jumla	jumla	jumlal
JAKARTA SELATAN	KEBAYORAN LAMA	GROGOL SELATAN	50	2	0	0	1
JAKARTA SELATAN	KEBAYORAN LAMA	KEBAYORAN LAMA SELAT	20	1	0	0	1
JAKARTA SELATAN	MAMPANG PRAPATAN	PELA MAMPANG	32.5	1	0	0	1
JAKARTA SELATAN	MAMPANG PRAPATAN	TEGAL PARANG	40	1	0	0	1
JAKARTA SELATAN	MAMPANG PRAPATAN	KUNINGAN BARAT	45	3	0	0	1
JAKARTA SELATAN	PESANGGRAHAN	ULUJAMI	30	1	0	0	1
JAKARTA SELATAN	TEBET	BUKIT DURI	40	2	0	0	1
JAKARTA SELATAN	TEBET	MANGGARAI	50	4	0	0	1
JAKARTA SELATAN	TEBET	KEBON BARU	50	1	0	0	1
JAKARTA SELATAN	PANCORAN	RAWAJATI	70	1	0	0	1
JAKARTA SELATAN	PANCORAN	PENGADEGAN	30	1	0	0	1
JAKARTA TIMUR	KRAMAT JATI	KAMPUNG TENGAH	30	1	0	0	1
JAKARTA TIMUR	KRAMAT JATI	CILILITAN	95	1	0	0	1
JAKARTA TIMUR	KRAMAT JATI	CAWANG	117.5	5	0	0	1
JAKARTA TIMUR	KRAMAT JATI	DUKUH	45	1	0	0	1
JAKARTA TIMUR	KRAMAT JATI	BALEKAMBANG	100	1	0	0	1
JAKARTA TIMUR	JATINEGARA	BIDARA CINA	100	2	0	0	1

Sumber: Data Kejadian Bencana Banjir Tahun 2024

Gambar 2. Dataset kejadian Banjir di Wilayah Jakarta

2. Tahap *Preprocessing*

Tahap *preprocessing* merupakan langkah krusial dalam proses analisis data untuk memastikan bahwa data yang digunakan dalam algoritma K-Means Clustering dan Naïve Bayes Classifier bersih, konsisten, dan siap untuk dianalisis (Zai, 2022). *Preprocessing* data dalam penelitian ini melibatkan beberapa langkah penting yang meliputi pembersihan data, penanganan data hilang, normalisasi data, dan pemisahan data untuk pelatihan dan pengujian (Chikalkar, 2020).

1. Pembersihan Data

   Langkah pertama dalam preprocessing adalah pembersihan data, yang bertujuan untuk mengidentifikasi dan memperbaiki atau menghapus data yang tidak valid atau anomali. Dalam dataset yang diperoleh dari Satu Data Jakarta, penulis memeriksa keberadaan duplikasi data, inkonsistensi dalam penamaan wilayah, kecamatan, dan kelurahan, serta kesalahan penulisan. Semua data yang tidak sesuai standar atau tidak relevan dihapus atau dikoreksi untuk memastikan kualitas data yang baik.

wilayah	kecamatan	kelurahan	jumlah_rata_rata_k	jumlah	jumlah	jumlal	jumlah_ke
JAKARTA SELATA	KEBAYORAN LAMA	GROGOL SELATAN	50	2	0	0	1
JAKARTA SELATA	KEBAYORAN LAMA	KEBAYORAN LAMA SE	20	1	0	0	1
JAKARTA SELATA	MAMPANG PRAPATA	PELA MAMPANG	32.5	1	0	0	1
JAKARTA SELATA	MAMPANG PRAPATA	TEGAL PARANG	40	1	0	0	1
JAKARTA SELATA	MAMPANG PRAPATA	KUNINGAN BARAT	45	3	0	0	1
JAKARTA SELATA	PESANGGRAHAN	ULUJAMI	30	1	0	0	1
JAKARTA SELATA	TEBET	BUKIT DURI	40	2	0	0	1
JAKARTA SELATA	TEBET	MANGGARAI	50	4	0	0	1
JAKARTA SELATA	TEBET	KEBON BARU	50	1	0	0	1
JAKARTA SELATA	PANCORAN	RAWAJATI	70	1	0	0	1
JAKARTA SELATA	PANCORAN	PENGADEGAN	30	1	0	0	1
JAKARTA TIMUR	KRAMAT JATI	KAMPUNG TENGAH	30	1	0	0	1
JAKARTA TIMUR	KRAMAT JATI	CILILITAN	95	1	0	0	1
JAKARTA TIMUR	KRAMAT JATI	CAWANG	117.5	5	0	0	1

Sumber: Hasil Penelitian (2024)

Gambar 3. Data Setelah Dibersihkan

Setelah melalui proses pembersihan, penulis mengidentifikasi, memperbaiki, dan menghapus beberapa data. Semula, data tersebut memiliki 15 atribut, yaitu: triwulan, bulan, wilayah, kecamatan, kelurahan, jumlah rata-rata ketinggian air, jumlah RW terdampak, jumlah KK terdampak, jumlah jiwa terdampak, jumlah kejadian, jumlah korban meninggal, jumlah korban luka, jumlah pengungsi, jumlah tempat pengungsian, dan nilai kerugian. Setelah dibersihkan, data tersebut disederhanakan menjadi 8 atribut, yaitu: wilayah, kecamatan, kelurahan, jumlah rata-rata ketinggian air, jumlah RW terdampak, jumlah KK terdampak, jumlah jiwa terdampak, dan jumlah kejadian.

2. Normalisasi data

   Normalisasi data dilakukan untuk memastikan bahwa semua atribut memiliki skala yang serupa dan menghindari dominasi atribut tertentu dalam analisis (Sirichanya & Kraisak, 2021). Dalam penelitian ini, atribut numerik seperti 'jumlah rata-rata ketinggian air', 'jumlah RW terdampak', 'jumlah KK terdampak', dan 'jumlah jiwa terdampak' dinormalisasi menggunakan metode *Min-Max Scaling*, yang mengubah nilai atribut ke dalam rentang [0, 1]. Proses normalisasi membantu dalam meningkatkan kinerja algoritma *K-Means Clustering* dan *Naïve Bayes Classifier.*

3. Pemisahan Data untuk Pelatihan dan Pengujian

Setelah proses pembersihan dan normalisasi selesai, data dibagi menjadi dua subset: data untuk pelatihan dan data untuk pengujian. Data pelatihan digunakan untuk mengembangkan model algoritma, sementara data pengujian berfungsi untuk menilai performa model tersebut. Pembagian data ini dilakukan dengan rasio 70%:30%, 80%:20%, dan 90%:10%.

3. Implementasi Gabungan Algoritma K-Means dan Naive Bayes dengan Rapidminer

Setelah pemahaman yang lebih dalam tentang data tercapai, penelitian melanjutkan dengan implementasi algoritma *K-Means*. Algoritma ini digunakan untuk mengelompokkan wilayah-wilayah berdasarkan karakteristik yang serupa terkait risiko banjir. Penggunaan algoritma ini bertujuan untuk mengidentifikasi pola-pola spasial yang mungkin tidak terlihat secara langsung dari data mentah.

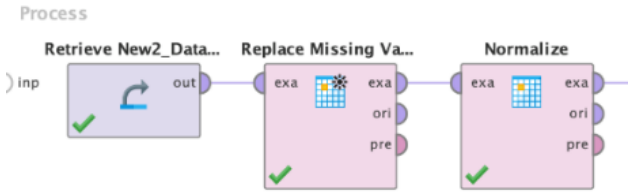
Sebelum melakukan pengolahan data di rapidminer perlu untuk *import dataset* yang akan digunakan, selanjutnya pastikan tipe datanya sudah sesuai dengan masing-masing atribut. Selanjutnya *drag and drop dataset* yang akan diproses.

Row No.	wilayah	kecamatan	kelurahan	jumlah_rat...	jumlah_rw_...	jumlah_kk_...	jumlah_jiw...	jumlah_kej...
1	JAKARTA SE...	KEBAYORAN...	GROGOL SE...	50	2	0	0	1
2	JAKARTA SE...	KEBAYORAN...	KEBAYORAN...	20	1	0	0	1
3	JAKARTA SE...	MAMPANG P...	PELA MAMP...	32.500	1	0	0	1
4	JAKARTA SE...	MAMPANG P...	TEGAL PAR...	40	1	0	0	1
5	JAKARTA SE...	MAMPANG P...	KUNINGAN ...	45	3	0	0	1
6	JAKARTA SE...	PESANGGRA...	ULUJAMI	30	1	0	0	1
7	JAKARTA SE...	TEBET	BUKIT DURI	40	2	0	0	1
8	JAKARTA SE...	TEBET	MANGGARAI	50	4	0	0	1
9	JAKARTA SE...	TEBET	KEBON BARU	50	1	0	0	1
10	JAKARTA SE...	PANCORAN	RAWAJATI	70	1	0	0	1
11	JAKARTA SE...	PANCORAN	PENGADEGAN	30	1	0	0	1
12	JAKARTA TI...	KRAMAT JATI	KAMPUNG T...	30	1	0	0	1
13	JAKARTA TI...	KRAMAT JATI	CILILITAN	95	1	0	0	1

ExampleSet (153 examples,0 special attributes,8 regular attributes)

Sumber: Hasil Penelitian (2024)
Gambar 4. *Dataset* yang telah diimport pada rapidminer

Selanjutnya akan menggunakan operator *replace missing value* untuk menghilangkan missing value dan perlu untuk menambahkan operator *Normalize* untuk menormalisasikan *dataset* yang sebelumnya diimport sebagai berikut.



Sumber: Hasil Penelitian (2024)
Gambar 5. Menambahkan Operator *Normalize*

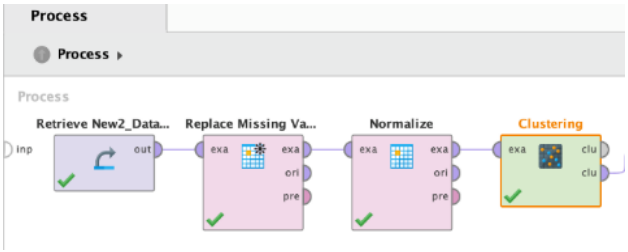
Diperoleh *output* sebagai berikut.

Row No.	jumlah_rat...	jumlah_rw_...	jumlah_kk_...	jumlah_jiw...	jumlah_kej...	wilayah	kecamatan	kelurahan
1	0.206	0.502	-0.154	-0.138	0.221	JAKARTA SE...	KEBAYORAN...	GROGOL SE...
2	-0.968	-0.412	-0.154	-0.138	0.221	JAKARTA SE...	KEBAYORAN...	KEBAYORAN...
3	-0.479	-0.412	-0.154	-0.138	0.221	JAKARTA SE...	MAMPANG P...	PELA MAMP...
4	-0.185	-0.412	-0.154	-0.138	0.221	JAKARTA SE...	MAMPANG P...	TEGAL PAR...
5	0.010	1.416	-0.154	-0.138	0.221	JAKARTA SE...	MAMPANG P...	KUNINGAN ...
6	-0.576	-0.412	-0.154	-0.138	0.221	JAKARTA SE...	PESANGGRA...	ULUJAMI
7	-0.185	0.502	-0.154	-0.138	0.221	JAKARTA SE...	TEBET	BUKIT DURI
8	0.206	2.330	-0.154	-0.138	0.221	JAKARTA SE...	TEBET	MANGGARAI
9	0.206	-0.412	-0.154	-0.138	0.221	JAKARTA SE...	TEBET	KEBON BARU
10	0.988	-0.412	-0.154	-0.138	0.221	JAKARTA SE...	PANCORAN	RAWAJATI
11	-0.576	-0.412	-0.154	-0.138	0.221	JAKARTA SE...	PANCORAN	PENGADEGAN
12	-0.576	-0.412	-0.154	-0.138	0.221	JAKARTA TI...	KRAMAT JATI	KAMPUNG T...
13	1.966	-0.412	-0.154	-0.138	0.221	JAKARTA TI...	KRAMAT JATI	CILILITAN

ExampleSet (153 examples,0 special attributes,8 regular attributes)

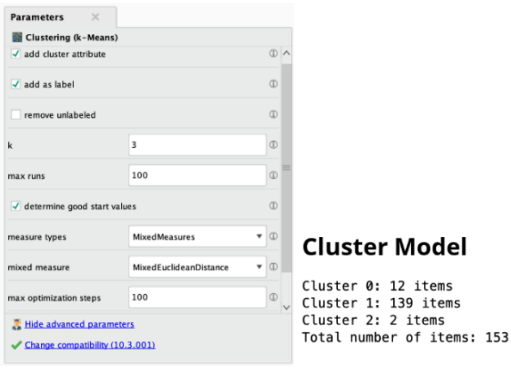
Sumber: Hasil Penelitian (2024)
Gambar 6. Hasil Normalisasi Data

Berikutnya menambahkan operator *K-Means Clustering* untuk mengklasterisasikan data kejadian bajir di DKI Jakarta sebagai berikut.



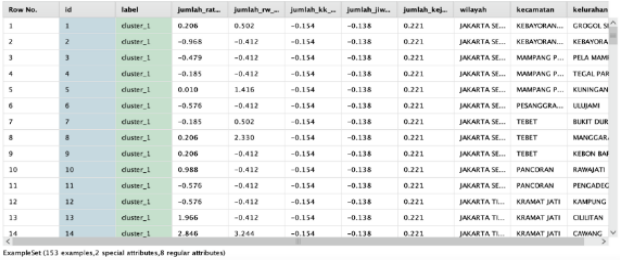
Sumber: Hasil Penelitian (2024)
Gambar 7. Penambahan Operator *K-Means Clustering*

Adapun pengaturan parameter pada operator *K-Means Clustering* menggunakan k=3 yakni membagi dataset menjadi 3 kategori yakni tinggi, rendah, sedang (Nigam & Rajavat, 2020). Pengkategorian tersebut didasarkan pada ketinggian air sebagai indikator utama. Kategori 'tinggi' menunjukkan area yang memerlukan prioritas evakuasi segera, sementara kategori 'sedang' mengindikasikan kebutuhan untuk bersiap-siap menghadapi potensi evakuasi. Kategori 'rendah' menunjukkan area yang belum memerlukan tindakan evakuasi dalam waktu dekat, namun tetap perlu dipantau. Selanjutnya *max runs*=100 yakni maksimal melakukan iterasi sebanyak 100 kali. *Measure types* yang digunakan adalah *mixed Measured* karena jenis tipe data yang ada pada tiap atribut beragam sehingga digunakan pengukuran campur. *Mixed measure* yang digunakan adalah *Mixed Euclidean Distance*, dengan pengaturan parameter sebagai berikut.



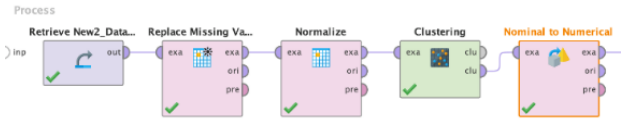
Sumber: Hasil Penelitian (2024)
Gambar 8. Setting Parameter pada operator *K-Means Clustering* dan Hasil Cluster Model K-Means

Berdasarkan hasil clustering untuk pengelompokkan data banjir berdasarkan pengkategorian ketinggian air sebagai indikator utama diperoleh *model cluster* 0 yang berarti "sedang" sebanyak 12 *items*, *cluster* 1 yang berarti "tinggi" sebanyak 139 *items,* dan juga cluster 2 yang berarti "rendah" diperoleh sebanyak 2 *items* dengan total *items* sebanyak 153.

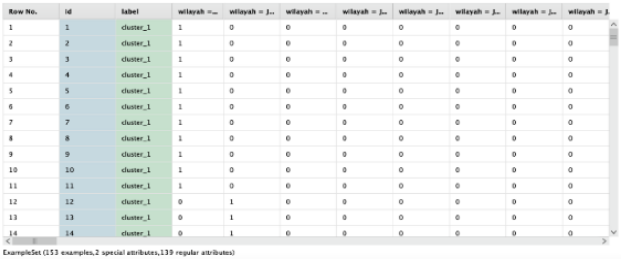


Sumber: Hasil Penelitian (2024)
Gambar 9. Dataset Hasil *Clustering K-Means*

Selanjutnya perlu untuk menambahkan operator *Nominal to Numerical* untuk mengubah atribut yang bernilai nominal menjadi bernilai numerical sebagai berikut.



Sumber: Hasil Penelitian (2024)
Gambar 10. Penambahan Operator *Nominal to Numerical*

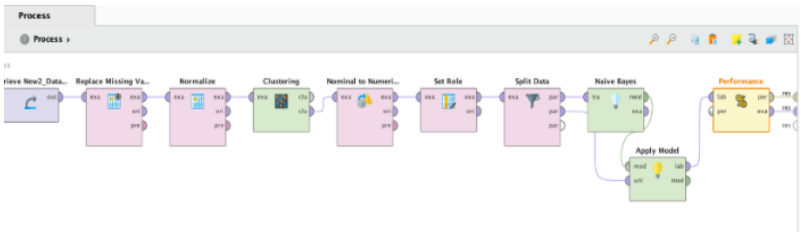


Sumber: Hasil Penelitian (2024)
Gambar 11. Dataset Hasil Penambahan Operator *Nominal To Numerical*

Sebelum melakukan klasifikasi *naïve bayes* maka perlu menambahkan operator *set role* untuk mengatur atribut label sebagai label.

Sumber: Hasil Penelitian (2024)
Gambar 12. Penambahan operator Set Role

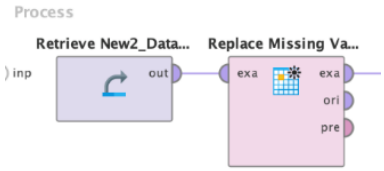
Langkah berikutnya adalah menambahkan operator pembagi data untuk memisahkan data menjadi dua subset, yaitu data pelatihan dan data pengujian. Data pelatihan digunakan untuk mengembangkan model algoritma, sedangkan data pengujian berfungsi untuk menilai performa model. Pembagian data ini dilakukan dalam tiga skenario proporsi, yaitu 70% data untuk pelatihan dan 30% untuk pengujian, 80% untuk pelatihan dan 20% untuk pengujian, serta 90% untuk pelatihan dan 10% untuk pengujian. Berikutnya menambahkan operator *Naïve Bayes*, *Apply Model* dan *performance* untuk mengklasifikasikan *dataset* pada data latih dan juga data uji sebagai berikut :



Sumber: Hasil Penelitian (2024)
Gambar 13. Proses Keseluruhan *K-Means Clustering* dan *Naïve Bayes*

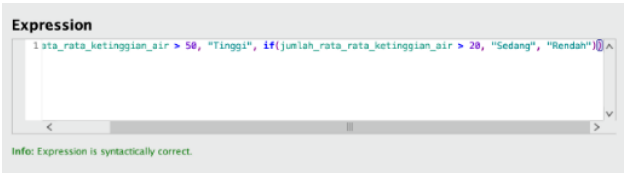
4. Implementasi Algoritma *Naïve Bayes* Konvensional dengan Rapidminer

Selain *K-Means*, penelitian juga melibatkan implementasi algoritma *Naïve Bayes*. Algoritma ini digunakan untuk membangun model klasifikasi yang dapat memprediksi wilayah-wilayah yang rentan terhadap banjir berdasarkan atribut-atribut yang diberikan. *Naïve Bayes* digunakan untuk memahami faktor-faktor yang berkontribusi terhadap risiko banjir dan memperkirakan probabilitas kejadian banjir di wilayah-wilayah tertentu (Erick et al., 2024). Sebelum melakukan pemrosesan data maka perlu untuk import dataset yang akan digunakan. Kemudian *drag and drop* pada *process* sebagai berikut.

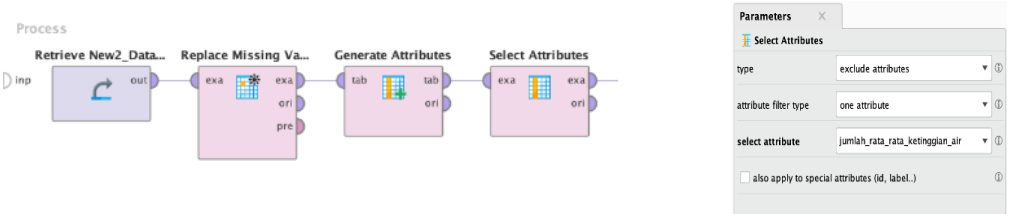


Sumber: Hasil Penelitian (2024)
Gambar 14. Penambahan Operator *Replace Missing Value*

Selanjutnya menambahkan operator *replace missing value* untuk menghilangkan *missing value*. Kemudian perlu untuk menambahkan operator *generate attributes* serta menambahkan *expression*. Selain itu perlu menambahkan operator *select attribute* serta *exclude attribute* jumlah_rata_rata_ketinggian_air sebagai berikut.

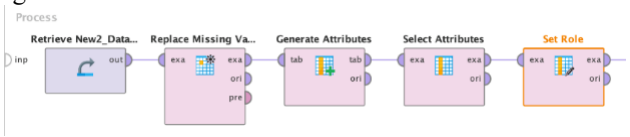


Sumber: Hasil Penelitian (2024)
Gambar 15. Penambahan Operator *Generate Attribute*



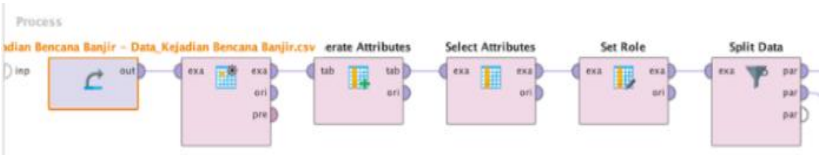
Sumber: Hasil Penelitian (2024)
Gambar 16. Penambahan Operator *Select Attribute*

Selanjutnya menambahkan operator *set role* untuk *setting attribute* label sebagai label untuk pemrosesan klasifikasi *Naïve Bayes* sebagai berikut.



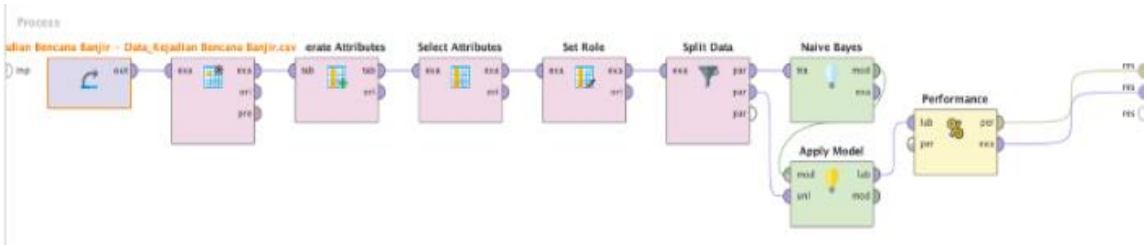
Sumber: Hasil Penelitian (2024)
Gambar 17. Penambahan Operator Set Role

Selanjutnya menambahkan operator *set role* untuk *setting attribute* label sebagai label untuk pemrosesan klasifikasi *Naïve Bayes* Kemudian menambahkan operator *Split Data* untuk membagi data menjadi 2 sub bagian yakni data uji dan dan data latih yang terbagi menjadi 3 skenario proporsi yakni 70% untuk pelatihan dan 30% untuk pengujian, 80% untuk pelatihan dan 20% untuk pengujian, 90% untuk pelatihan dan 10% untuk pengujian. Hal tersebut dilakukan untuk memastikan bahwa model yang dihasilkan memiliki generalisasi yang baik terhadap data baru sebagai berikut.



Sumber: Hasil Penelitian (2024)
Gambar 18. Penambahan Operator *Split Data* dan *Setting Split Data*

Langkah terakhir adalah menambahkan operator performance untuk mengetahui kinerja naïve bayes dalam mengklasifikasikan *dataset* daerah rawan banjir Kota Jakarta tersebut.



Sumber: Hasil Penelitian (2024)
Gambar 19. Proses Lengkap Klasifikasi *Naïve Bayes* Konvensional

5. Evaluasi dan Validasi

Tahap berikutnya meliputi evaluasi dan validasi model yang dikembangkan dengan menggunakan algoritma *K-Means* dan *Naïve Bayes*. Evaluasi bertujuan untuk menilai performa model dalam mengklasifikasikan daerah-daerah yang rawan banjir. Metrik evaluasi yang digunakan mencakup akurasi dan *weighted mean recall* (Riyanto et al., 2023). karena dalam banyak kasus, dataset memiliki kelas yang tidak seimbang (*imbalanced class*). *Weighted mean recall* memberikan bobot yang berbeda pada setiap kelas berdasarkan jumlah sampelnya sementara Dalam konteks integrasi *K-Means* dan *Naive Bayes*, akurasi penting untuk melihat seberapa baik model mengklasifikasikan data setelah proses *clustering*.Validasi dilakukan untuk memastikan bahwa model yang dibangun dapat digeneralisasi dengan baik pada data baru yang belum pernah dilihat sebelumnya, hasil *performance* dapat dilihat pada tabel berikut:

Tabel 1. Hasil Perbandingan *Performance*

	Akurasi			Weighted Mean Recall			Mean	
	70:30	80:20	90:10	70:30	80:20	90:10	Akurasi	WMR
Penggabungan K-Means dan Naïve Bayes	97,87%	96.67%	100%	66.67%	65.48%	66.67%	98.18%	66.27%
Naïve Bayes Konvensional	40,43%	43.33%	46.67%	56.04%	50.44%	36.67%	43.47%	47.72%

Sumber: Hasil Penelitian (2024)

Berdasarkan Tabel 1, terlihat bahwa metode "Penggabungan *K-Means* dan *Naïve Bayes*" mencapai tingkat akurasi tertinggi, yaitu 98.18%. Hasil ini merupakan yang tertinggi dibandingkan dengan metode "*Naïve Bayes* Konvensional" yang hanya mencapai akurasi 43.47%. Perbandingan ini menunjukkan bahwa penggabungan metode *K-Means* dan *Naïve Bayes* secara signifikan meningkatkan akurasi dalam klasifikasi data. Dengan demikian, dapat disimpulkan bahwa metode "Penggabungan *K-Means* dan *Naïve Bayes*" sangat efektif dalam meningkatkan akurasi klasifikasi. Hasil akurasi sebesar 98.18% menunjukkan bahwa metode ini memiliki kinerja yang sangat baik dan dapat diandalkan untuk tugas-tugas klasifikasi data yang kompleks. Penelitian ini memberikan bukti kuat tentang potensi penggabungan metode dalam meningkatkan kinerja data mining.

KESIMPULAN

Hasil dari penelitian ini menunjukkan bahwa penggabungan algoritma *K-Means Clustering* dan *Naïve Bayes Classifie*r memiliki performa yang lebih unggul dibandingkan dengan penggunaan algoritma *Naïve Bayes* konvensional dalam mengklasifikasikan wilayah rawan banjir di kota Jakarta. Dalam hal rata-rata akurasi, model yang menggunakan kombinasi kedua algoritma ini berhasil mencapai tingkat rata-rata akurasi sebesar 98.18% pada rasio split data 70:30, 80:20, 90:10 yang menunjukkan kemampuan model dalam mengklasifikasikan data dengan benar. Selain itu, nilai *Weighted Mean Recall* yang diperoleh sebesar 66.67% menunjukkan bahwa model memiliki tingkat keberhasilan yang baik dalam mendeteksi kelas-kelas banjir yang lebih jarang terjadi.

REFERENSI

Alghifari, F., & Juardi, D. (2021). *Fauzan Alghifari Penerapan Data Mining Pada Penerapan Data Mining Pada Penjualan Makanan Dan Minuman Menggunakan Metode Algoritma Naïve Bayes*.

Anggraini, N., Pangaribuan, B., Siregar, A. P., Sintampalam, G., Muhammad, A., Ridha, M., Damanik, S., & Rahmadi, T. (2021). *ANALISIS PEMETAAN DAERAH RAWAN BANJIR DI KOTA MEDAN TAHUN 2020*. *4*(2).

Angreini, S., & Supratman, E. (2021). Visualisasi Data Lokasi Rawan Bencana Di Provinsi Sumatera Selatan Menggunakan Tableau. *Jurnal Nasional Ilmu Komputer* , *2*.

Badan Nasional Penanggulangan Bencana. (2023). *BUKU IRBI 2023*.

Bui, M. A., & Bahtiar, A. (2024). IMPLEMENTASI METODE ALGORITMA K-MEANS CLUSTERING UNTUK MENGELOMPOKKAN TRANSAKSI PENJUALAN BARANG DI TOKO ARINO. In *Jurnal Mahasiswa Teknik Informatika* (Vol. 8, Issue 2).

Chikalkar, S. N. (2020). Knowledge Discovery and Data Mining. *International Journal for Research in Applied Science and Engineering Technology*, *8*(10), 874–876. https://doi.org/10.22214/ijraset.2020.32045

Effendi, M. M., Inka, I., & Siswandi, A. (2024). Analysis Prediksi Wilayah Rawan Banjir dengan Algoritma K-Means. *Journal of Information System Research (JOSH)*, *5*(2), 697–703. https://doi.org/10.47065/josh.v5i2.4770

Erick, B., Akhmad, F., & Ibnu Prasetyo, W. (2024). Application of Naive Bayes Algorithm for Physical Fitness Level Classification. *International Journal of Disabilities Sports and Health Sciences*, *7*(1), 178–187. https://doi.org/10.33438/ijdshs.1330745

Fatonah, N. S., Buana, M., Selatan, J. M., Kembangan, K., Barat, J., Khusus, D., Jakarta, I., & Com, N. (2021). *Penerapan Deteksi Bencana Banjir Menggunakan Metode Machine Learning*.

Khomsiyah, J., Ramdhani, A., Damayanti, A. F., & Rohman, D. (2021). *PENERAPAN ALGORITMA K-MEANS CLUSTERING UNTUK PENGELOMPOKAN WILAYAH RAWAN BANJIR*.

Martin Saputra. (2025). *FLOOD PREDICTION WITH NAÏVE BAYES METHOD*.

Nandang Iriadi, Priatno, & Ahmad Ishaq. (2020). *Penerapan Data Mining dengan Rapid Miner; Konsep Data Mining, Data Warehouse, Metode, Model, Teknik*. Graha Ilmu.

Nigam, N., & Rajavat, A. (2020). A Systematic Literature Review of Data Classification Techniques. In *International Journal of Computer Applications* (Vol. 177, Issue 44). https://www.sas.com/en_us/insights/analytics/data-

Ridwan, A. (2020). *Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus*.

Riyanto, S., Sitanggang, I. S., Djatna, T., & Atikah, T. D. (2023). Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 14, Issue 6). http://gcancer.org/pdr

Sirichanya, C., & Kraisak, K. (2021). Semantic data mining in the information age: A systematic review. *International Journal of Intelligent Systems*, *36*(8), 3880–3916. https://doi.org/https://doi.org/10.1002/int.22443

Zai, C. (2022). IMPLEMENTASI DATA MINING SEBAGAI PENGOLAHAN DATA. In *Portaldata.org* (Vol. 2, Issue 3).

Zhang, X. (2020). Research on Data Mining Algorithm Based on Pattern Recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, *34*(6). https://doi.org/10.1142/S0218001420590156