Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 219−228
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5958

Analisis Prediksi Banjir di Indonesia Menggunakan Algoritma Support Vector Machine dan Random Forest

Indarto Aditya Purnomo*, Jamaludin Indra, Elsa Elvira Awal, Tatang Rohana

Teknik Informatika, Universitas Buana Perjuangan Karawang, Karawang
Jl. HS.Ronggo Waluyo, Puseurjaya, Telukjambe Timur, Karawang, Jawa Barat, Indonesia
Email: 1,*if20.indartopurnomo@mhs.ubpkarawang.ac.id, 2jamaludin.indra@ubpkarawang.ac.id
3elsaelvira@ubpkarawang.ac.id, 4tatang.rohana@ubpkarawang.ac.id
Email Penulis Korespondensi: if20.indartopurnomo@mhs.ubpkarawang.ac.id
Submitted: **24/09/2024;** Accepted: **07/10/2024;** Published: **15/10/2024**

Abstrak−Bencana alam sering terjadi di indonesia seperti banjir, tanah longsor, dan gunung meletus. Faktor geologis, seperti pertemuan empat lempeng utama, membuat Indonesia rentan terhadap bencana alam. Data statistik dari Badan Penanggulangan Bencana Alam menunjukkan peningkatan kejadian banjir setiap tahun, mencapai puncaknya pada tahun 2021 dengan 1.794 kejadian. Antisipasi diperlukan sejak dini untuk meminimalisir dampak bencana alam, dan pola prediksi menjadi pengetahuan baru untuk mencegah dan menanggulangi bencana tersebut. Pada penelitian ini menerapkan algoritma Support Vector Machine dan Random Forest. Hasil dari penelitian ini prediksi banjir pada tahun 2024 – 2026 terbesar di Indonesia ialah Aceh sebanyak 240 banjir, Sumatera Utara 215 banjir, Jawa Barat sebanyak 210 banjir dan Jawa Tengah sebanyak 160 banjir, hasil perbandingan algoritma terbaik di peroleh Random Forest sebesar 99,6% dengan nilai rata – rata RSME 3.853.

Kata Kunci: Banjir; Support Vector Machine; Random Forest; RSME; Prediksi

Abstract−Natural disasters frequently occur in Indonesia, such as floods, landslides, and volcanic eruptions. Geological factors, such as the convergence of four major tectonic plates, make Indonesia vulnerable to natural disasters. Statistical data from the National Disaster Management Agency show an increase in flood occurrences each year, peaking in 2021 with 1,794 incidents. Early anticipation is necessary to minimize the impact of natural disasters, and predictive patterns are becoming new knowledge for preventing and managing these disasters. This study applies the Support Vector Machine and Random Forest algorithms. The results of this study predict that the largest number of floods from 2024 to 2026 in Indonesia will occur in Aceh with 240 floods, North Sumatra with 215 floods, West Java with 210 floods, and Central Java with 160 floods. The best algorithm comparison results were achieved with Random Forest, which had an accuracy of 99.6% and an average RMSE value of 3.834.

Keywords: Flood; Support Vector Machine; Random Forest; RSME; Prediction

1. PENDAHULUAN

Indonesia merupakan negara yang sangat beresiko akan datangnya sebuah bencana alam yang relatif sangat tinggi, bencana alamnya yang sering terjadi yaitu tsunami, banjir, kebakaran hutan, tanah longsor dan gunung Meletus[1]. Secara geologis, Indonesia sering menimbulkan kejadian bencana alam di beberapa daerah, seperti di Jakarta yang setiap tahunnya terjadi Banjir [2]. Dikarenakan Indonesia berada diposisi pertemuan 4 benua utama yang meliputi Euroasia, Indo – Autralia, Filipina dan Pasifik yang mengakibatkan Indonesia relatif rawan bencana alam[3]. Bencana Alam ialah fenomena alam yang terjadi tanpa kita sadari, mengacu pada UU No 24 tahun 2007 yang didefinisikan sebuah kejadian atau peristiwa yang menyebabkan kehidupan manusia menjadi terganggu, yang berasal dari sebuah kondisi alam atau aspek dari kelalaian manusia, mengakibatkan kematian, kehilangan harta benda, kerusakan lingkungan dan efek psikologi [4], [5].

Menurut data yang telah dikumpulkan dari situs web Badan Nasional Penanggulangan Bencana, pada bencana banjir setiap tahun mengalami baik peningkatan maupun penurunan, seperti yang terlihat pada tahun 2015 mengalami kejadian sebanyak 525 kejadian, dan pada tahun 2021 mengalami peningkatan kejadian banjir sebanyak 1,794 kejadian [6]. Dari data yang telah diperoleh untuk mengantisipasi sejak awal perlu dilaksanakan agar bisa meminimalisir terjadi dampak yang akan menimbulkan suatu bencana banjir sehingga diperlukannya suatu rangkaian prediksi sebagai wawasan baru untuk menanggulangi terjadinya bencana banjir. Berdasarkan data dari Geoportal Bencana Indonesia dan Badan Nasional Penanggulangan Bencana (BNPB) dalam laman website memperlihatkan data yang signifikan dari kejadian banjir di Indonesia[7]. Data yang telah diambil pada tanggal 16 Januari 2024, Tercatat pada awal Januari 2019 sampai Desember 2023 telah terjadi bencana banjir sebanyak 7,168 Kejadian banjir di Indonesia [8]. Dari data yang telah dikumpulkan menimbulkan korban jiwa maupun korban luka – luka, yang menghancurkan tempat kejadian berupa infrastruktur, kerugian material dan immaterial [9].

Data mining mempunyai salah satu algoritma ialah Support Vector Machine yang mempunyai suatu kelebihan untuk menyelesaikan sebuah masalah prediksi dengan akurasi yang terbaik dibandingkan dengan Naïve Bayes [10]. Data mining menggunakan algoritma seperti Random Forest, mempunyai kelebihan dalam membuat suatu prediksi dalam memproses akurasi terbaik[11].

Penelitian tentang Prediksi Banjir untuk penentuan Daerah Rawan banjir dengan Algoritma Support Vector Machine Dilakukan oleh Dwiasnati dan Devianto menggunakan metode algoritma SVM, yang menghasilkan tingkat akurasi sebesar 85,71% dan UAC menghasilkan sebesar 0.481 sedangkan nilai akurasi menggunakan

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 219−228
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5958

algoritma support Vector machine berbasis PSO sebesar 97,62 % dan AUC yang menghasilkan nilai 1.000 [12]. Penelitian tentang komparasi Algoritma Klasifikasi C5.0, Support Vector Machine dan Naive Bayes untuk Prediksi Banjir yang dilakukan oleh Fitrianah, Gunawan dan Sari dengan metode Algoritma C5.0, SVM dan Naïve Bayes, yang menghasilkan tingkat nilai akurasi C5.0 dan SVM 93,75 %, dan pada metode algoritma Naïve Bayes akurasi 81,25% [13]. Penelitian tentang Analisa Prediksi Angin Puyuh di Jawa Barat menggunakan metode Algoritma K-Nearest Neighbor dan C4.5 yang dilakukan oleh Nuryaman, Yudha, dan Asistyasari memakai metode algoritma K-Nearest Neighbor berbasis PSO mendapatkan hasil akurasi sebesar 83%, presisi 85% recall 96,15%, dan AUC 0,591 [5] Penelitian tentang Analisa Algoritma Decision Tree, Random Forest, dan Naïve Bayes untuk Prediksi Banjir di Desa Dayeuh kolot yang dilakukan oleh Darmawan, Dewanta dan Astuti mendapatkan hasil algoritma Random Forest memiliki hasil terbaik sebesar 97.40 dibandingkan Decesion Tree dan Naive Bayes [14].

Menurut penjelasan penelitian yang telah dilakukan dan hasil penelitian, maka ada kemungkinan bahwa prediksi kasus banjir dapat digunakan, untuk memprediksi kejadian banjir yang ada di Indonesia menggunakan metode algoritma Support Vector Machine dan Random Forest untuk memperkirakan kejadian pada tahun yang akan mendatang, dan didukung oleh data peristiwa sebelumnya pada tahun 2019 – 2023 dari Geoportal Bencana Indonesia (BNPB). Hasil dari sebuah penelitian ini diharapkan dapat membantu Pemerintah terkait dan yang terpenting Masyarakat umum Indonesia dalam mengantisipasi terjadinya bencana alam tahun depan.

2. METODOLOGI PENELITIAN

2.1 Tahahapan Penelitian

Berikut gambar 1 merupakan tahapan dari penelitian yang dilakukan.

Gambar 1. Diagram Tahapan Penelitian

a. Pengumpulan Data ini dikumpulkan dari Geoportal Data Bencana Indonesia di laman web gis.bnpb.co.id memperlihatkan data sebanyak 7,168 Kejadian. Dari Tercatat pada awal Januari 2019 sampai Desember 2023.
b. Analisa data merupakan tahap Analisis agar data yang menjadi lebih mudah diproses ke tahap selanjutnya. Ada beberapa macam tahap Analisis data yaitu jumlah missing values, jumlah Duplikasi Data dan Jumlah persentase data yang adakan dihapus.
    1. Jumlah missing values Pada tahap ini, missing values dalam data mengacu pada nilai yang tidak ada atau kosong untuk suatu variabel tertentu. Jumlah missing values dapat bervariasi tergantung pada dataset dan variabel yang diamati.
    2. Jumlah duplikasi data pada tahap ini, setalah mencari missing value selanjutnya mencari jumlah kesamaan data menggantikan dengan nilai 0.
    3. Jumlah presentase data yang akan dihapus pada tahap ini, setelah mencari missing values dan mencari jumlah duplikasi data, selanjutnya data yang akan disaring pada tahap ini berguna untuk menghindari data yang sama dan nilai 0.
c. Preprocessing Data merupakan bagian dari tahap-tahap untuk mengolah data yang dapat diakses dan dipahami sehingga dapat digunakan untuk tahap selanjutnya dari proses. Adapun beberapa tahapan di Preprocessing Data yaitu Cleaning data, Transformasi data, Ekstraksi Fitur dan mencari Pola data per tahun.
    1. Cleaning data pada proses ini, melibatkan memperbaiki data yang tidak signifikan. Biasanya, data yang didapatkan dari berbagai sumber, termasuk dataset, yang berisikan data yang tidak sempurna sebaiknya dihilangkan agar performa teknik data mining tidak terpengaruh. Dengan membersihkan data, jumlah dan kompleksitas data yang ditangani akan berkurang.
    2. Transformasi Data pada tahap transformasi, pada dataset yang berisikan tipe data yang diproses yang semula data mentah menjadi data bentuk yang lebih dapat di interpretasikan atau lebih sesuai untuk analisa statitik seperti tipe data numerik dan x, y.
    3. Under Sampling pada tahap ini, proses untuk menangani sebuah ketidakseimbangan data pada suatu kelas dengan cara mengurangi jumlah sampel dari kelas mayoritas. Bertujuan untuk mencapai sebuah keseimbangan seluruh data.
    4. Ekstraksi Fitur pada tahap ini, proses mengidentifikasi, memilih dan mengubah data mentah menjadi representasi fitur yang lebih informatif untuk mengekstrak fitur – fitur dari data untuk memperbaiki kinerja model dan meningkatkan akurasi prediksi.
d. Processing Data merupakan serangkaian alur untuk proses menjadi sebuah informasi data. Adapun tahapan di Processing data yaitu Feature Selection, Mencari pola data pertahun dan Split Data.

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 219−228
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5958

1. Feature Selection pada tahap ini, merupakan proses pemilihan fitur yang relavan dan penting dalam pemodelan atau analisis data. Teknik ini bertujuan untuk mengurangi Jumlah fitur yang digunakan dalam pembelajaran, sambal memilih fitur yang memiliki kemampuan diskriminasi yang tinggi. Dengan memilih kombinasi fitur yang optimal dan seleksi fitur dapat meningkatkan akurasi model dan menggunakan visualisasi Heatmap.
2. Mencari pola data pertahun pada tahap ini, Proses analisis data untuk mengidentifikasi tren dan pola dalam data dari tahun ke tahun dalam data yang dikumpulkan pada waktu tertentu, dengan cara menggunakan visualisasi data yang menggunakan scatter plot dengan sumbu x yang mewakili tahun dan sumbu y mewakili nilai yang sedang diamati.
3. Split Data pada tahap ini, tindakan memisahkan dataset menjadi dua atau lebih bagian yang berbeda, masing-masing digunakan untuk keperluan tertentu seperti pelatihan, validasi, dan pengujian model.

e. Support Vector Machine (SVM) ialah algoritma klasifikasi linear dimana mencari hyperlane yang cocok digunakan untuk sebagai pemisah dua kelas pada ruang input [15]. Pengembangan klasifikasi linear agar mudah di proses pada permasalahan non linear[16]. Agar data pelatihan asli menjadi data pelatihan dengan dimensi yang lebih baik [17].

f. Random Forest salah satu algoritma untuk meningkatkan sebuah hasil akurasi terbaik dengan adanya pemilihan secara acak untuk setiap node [18] dan untuk mengklasifikasikan hasil dari pohon keputusan berdasarkan kategori yang paling sering digunakan [19].

g. Evaluasi RMSE Selanjutnya dalam tahap evaluasi dilakukan menggunakan Root Mean Square eror (RMSE) ialah dasar dari kuadrat kesalahan yang menghasilkan sebuah perhitungan untuk mencari nilai terendah agar mendapatkan hasil nilai aslinya[13]. Dengan cara menghitung Mean square eror , Mean absolute percentage eror dan Root Mean Square eror untuk mengukur tingkat akurasi model prediksi [20]

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\text{Yi}-\hat{\text{Y}}\text{i})2}{n}} \quad (1)$$

Dalam sebuah analisis data, γi merepresentasikan nilai data ke-i yang sebenarnya, sementara y^i adalah nilai estimasi dari data ke-i. Nilai n mengacu pada jumlah keseluruhan data yang digunakan dalam analisis.

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Hasil pembahasan, mengunakan dataset di peroleh dalam laman website gis.bnpb.go.id yang mempunyai keseluruhan data sebanyak 114.688. Data pada tahun 2014 – 2023, 16 kolom dan jumlah baris 7.168. Contoh 5 dataset teratas seperti pada tabel 1.

Tabel 1. Pengumpulan Data

No.	Kode Identitas Bencana	ID Kabupaten	Tanggal Kejadian	Kejadian	Lokasi
1	1613 101 2023 012 031 1	1613	12/31/2023	BANJIR	Kec. Rupit Ds. Noman Ds. Noman Baru Ds. Batu Gajah Ds.Batu Gajah Baru Ds. Maur Lama Ds. Maur Baru Ds.Tanjung Beringin Ds.Bingin Rupit Ds.Beringin Jaya Kec. Karang Jaya Ds.Ma Batang Empu Ds.Rantau Telang Ds.Suka Menang Ds.Bukit Langkap Ds.Muara Tiku Ds.Embacang Lama Ds.Bukit Ulu Ds.Rantau Jaya
2	1509 101 2023 012 031 1	1509	12/31/2023	BANJIR	Kec VII Koto Desa Teluk Lancang
3	3215 101 2023 012 031 1	3215	12/31/2023	BANJIR	Kec. Teluk Jambe Ds. Karang Liar
4	1307 101 2023 012 031 1	1307	12/31/2023	BANJIR	Kec.Pangkalan Nagari Pangkalan Nagari Koto Alam Ke Kapur IX Nagari Sialang Nagari Muaro Paiti Nagari Galugua Kec Lareh Sago Halaban Nagari Sitanang Kec Suliki Nagari kurai
5	3201 101 2023 012 031 1	3201	12/31/2023	BANJIR	Kec. Katapang Ds. Banyusari Kec. Tajurhalang

Kabupaten	Provinsi	Kronologi & Dokumentasi	Penyebab

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 219−228
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5958

MUSI RAWAS UTARA	SUMATERA SELATAN	Dokumentasi	Hujan dengan intensitas sedang - tinggi yang mengguyur wilayah kabupaten Musi Rawas Utara pada hari Sabtu - Minggu yang mengakibatkan sungai rupit meluap
TEBO	JAMBI	Dokumentasi	
KARAWANG	JAWA BARAT	Dokumentasi	Hujan deras di wilayah hulu dan beberapa wilayah kabupaten Karawang sehingga Meluapnya sungai Citarum dan Cibeet serta sungai Cidawolong
LIMA PULUH KOTA	SUMATERA BARAT	Dokumentasi	Pusdalops PB melakukan koordinasi dengan pihak Kecamatan dan Nagari , POLRI, TNI dan OPD terkait. TRC melakukan kaji cepat. Kondisi Mutakhir : Banjir sudah mulai surut Masyarakat rumah terdampak banjir dan longsor mengungsi ke rumah keluarga
BOGOR	JAWA BARAT	Dokumentasi	Dipicu oleh hujan dengan intensitas tinggi

Meninggal	Hilang	Terluka	Rumah Rusak	Rumah Terendam	Fasum Rusak
NaN	NaN	NaN	1	4031	1
NaN	NaN	NaN	0	44	0
NaN	NaN	NaN	0	662	10
NaN	NaN	NaN	2		0
NaN	NaN	NaN	0	660	1

3.2 Analisa Data

Pada tahap Analisa data terdapat mising value sebanyak 1.251 data dalam dataset, jumlah duplikasi data sebanyak 0 data duplikat dan pada persentase data yang akan dihapus sebanyak 2.18% dari total dataset dikarenakan ketidaklengkapan data tersebut terlihat pada kolom gambar 2.

```
Jumlah Missing Value: 1251
Jumlah Duplikasi Data: 0
Persentase Data yang akan dihapus: 2.1815708705357144%
```

Gambar 2. Analisa Data

3.3 Preprocessing Data

Setelah data terkumpulkan dan analisa data, pada proses ini terjadi menghapus missing value, mengganti data non numerik menjadi numerik, menghapus baris kosong, menghapus duplikasi data menghapus kolom kosong seperti pada gambar 3.

```
Jumlah Missing Value: 0
---------------
Jumlah Data yang bukan numeric: 0
---------------
Jumlah Duplikasi Data: 0
```

Gambar 3. Preprocessing Data

Setelah tahap sebelumnya, data yang diperlukan seperti tanggal kejadian, kejadian, provinsi dan penyebab. Selanjutnya, data tanggal kejadian diubah menjadi format tahun, bulan, dan tanggal untuk memudahkan proses pada tahap berikutnya, sesuai dengan contoh pada gambar 4.

```
DataFrame with extracted Year, Month, and Day:
  Tanggal Kejadian  Tahun  Bulan  Tanggal
0       2023-12-31   2023     12       31
1       2023-12-31   2023     12       31
2       2023-12-31   2023     12       31
3       2023-12-30   2023     12       30
4       2023-12-25   2023     12       25
```

Gambar 4. Mengubah data tanggal kejadian

Setelah tahap pengubahan tanggal kejadian pada dataset penyebab, untuk mempermudah analisis pada tahap selanjutnya, data tersebut dikelompokkan menjadi 6 kategori utama: hujan dan intensitas curah hujan, luapan sungai, saluran tersumbat, struktur tanah dan longsor, gelombang pasang, dan tidak ada penyebab. Langkah-langkah pengelompokan ini dilakukan sesuai dengan yang ditunjukkan pada tabel 2.

Tabel 2. Tahap Filter Data Penyebab

Penyebab	Kategori Penyebab	Numerik
	Tidak Ada Penyebab	0

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 219−228
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5958

Penyebab	Kategori Penyebab	Numerik
Penyebab yang terkait langsung dengan hujan dengan intensitas tinggi.	Hujan Lebat	1
Penyebab yang terkait dengan luapan sungai atau banjir akibat luapan sungai	Luapan Sungai	2
Penyebab yang melibatkan angin kencang	Angin Kencang	3
Penyebab yang terkait dengan drainase atau sistem pembuangan air yang buruk	Drainase Buruk	4
Banjir yang terjadi secara tiba-tiba karena aliran air yang deras	Banjir Bandang	5
Penyebab yang terkait dengan pergerakan tanah atau longsoran.	Tanah Longsor	6
Penyebab yang terkait dengan pasang naik air laut atau gelombang tinggi	Pasang Laut	7
Penyebab yang merupakan kombinasi dari beberapa faktor seperti hujan dan angin	Faktor Kombinasi	8
Penyebab yang kombinasi keseluruhan	Penyebab Lain	9

Setelah tahap pengelompokkan penyebab, data non-numerik seperti kolom kejadian dan kabupaten diubah menjadi data numerik menggunakan transformasi data. Selain itu, data yang masih berbentuk NaN (Not a Number) diubah menjadi data numerik, yang dapat dilihat di gambar 5.

```
DataFrame with NaN values replaced by mean:
   Kejadian  Provinsi                                          Penyebab  \
0         1        37  • Hujan dengan intensitas sedang - tinggi yang...
1         1         8                                               NaN
2         1         9  • Hujan deras di wilayah hulu dan beberapa wil...
3         1        36  • Pusdalops PB melakukan koordinasi dengan pih...
4         1         9       • Dipicu oleh hujan dengan intensitas tinggi

   Tahun  Bulan  Tanggal  Numerik Penyebab   Kategori Penyebab
0   2023     12       31                 2       Luapan Sungai
1   2023     12       31                 0  Tidak Ada Penyebab
2   2023     12       31                 2       Luapan Sungai
3   2023     12       31                 6       Tanah Longsor
4   2023     12       31                 1         Hujan Lebat
```

Gambar 5. Transformasi NaN

Setelah tahap penanganan nilai NaN (Not a Number), langkah selanjutnya adalah mengonversi nama provinsi di Indonesia menjadi format numerik, seperti yang ditunjukkan pada tabel 3.

Tabel 3. Provinsi Numerik

Provinsi	Numerik	Provinsi	Numerik	Provinsi	Numerik
Aceh	0	Jawa Tengah	10	Maluku	20
Bali	1	Jawa Timur	11	Maluku Utara	21
Banten	2	Kalimantan Barat	12	Nusa Tenggara Barat	22
Bengkulu	3	Kalimantan Selatan	13	Nusa Tenggara Timur	23
Yogyakarta	4	Kalimantan Tengah	14	P A P U A	24
DI Yogyakarta	5	Kalimantan Timur	15	Papua	25
Jakarta	6	Kalimantan Tenggara	16	Papua Barat	26
Gorontalo	7	Bangka	17	Papua Barat Daya	27
Jambi	8	Kepulauan Riau	18	Papua Pegunungan	28
Jawa Barat	9	Lampung	19	Papua Selatan	29

Provinsi	Numerik
Riau	30
Sulawesi Barat	31
Sulawesi Selatan	32
Sulawesi Tengah	33
Sulawesi Tenggara	34
Sulawesi Utara	35
Sumatera Barat	36
Sumatera Selatan	37
Sumatera Utara	38

Setelah menggantikan nilai **NaN** dan mengonversi kolom provinsi menjadi data numerik, Tabel 4 menunjukkan data jumlah kejadian bencana sebelum dilakukan undersampling. Dalam tabel tersebut, kejadian hujan lebat mendominasi dengan 2.248 kasus, disusul oleh luapan sungai sebanyak 1.815 kasus dan kejadian tanpa penyebab yang teridentifikasi sebanyak 1.550 kasus. Kategori-kategori bencana lainnya seperti banjir bandang,

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 219−228
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5958

tanah longsor, pasang laut, angin kencang, dan drainase buruk juga tercatat, namun dengan jumlah yang lebih rendah, misalnya banjir bandang sebanyak 664 kasus, dan penyebab lainnya dengan jumlah terendah sebanyak 38 kasus.

Tabel 4. Sebelum Undersampling

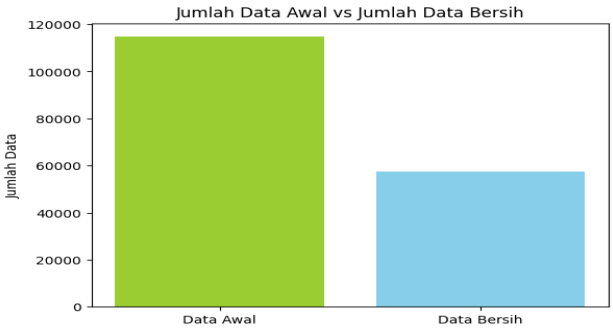
Hujan Lebat	Luapan Sungai	Tidak ada Penyebab	Banjir Bandang	Tanah Longsor	Pasang Laut	Angin Kencang	Drainase Buruk	Penyebab Lainnya	Faktor Kombinasi
2248	1815	1550	664	286	225	179	160	38	3

Setelah dilakukan undersampling, sebagaimana ditunjukkan pada Tabel 5, seluruh kategori bencana disesuaikan dengan jumlah data yang lebih seimbang, yaitu masing-masing hanya 3 kasus. Teknik undersampling ini dilakukan untuk memastikan bahwa setiap kategori bencana memiliki jumlah data yang proporsional, guna menghindari bias dalam model analisis yang akan digunakan.

Tabel 5. Sesudah Undersampling

Hujan Lebat	Luapan Sungai	Tidak ada Penyebab	Banjir Bandang	Tanah Longsor	Pasang Laut	Angin Kencang	Drainase Buruk	Penyebab Lainnya	Faktor Kombinasi
3	3	3	3	3	3	3	3	3	3

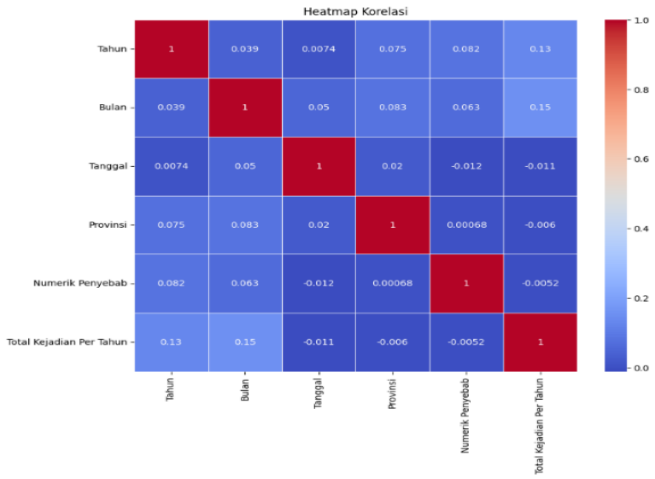
Setelah under sampling, data tersebut kemudian dibandingkan menggunakan barplot. Pada barplot tersebut, data awal yang berjumlah 114.688. ditampilkan dengan warna hijau, sementara data bersih yang berjumlah 57.344 ditampilkan dengan warna biru. Perbandingan ini bertujuan untuk mempermudah proses selanjutnya dan pada gambar 6.



Gambar 6. Data awal vs Data bersih.

3.4 Processing Data

Pada pembahasan ini, data telah diproses dipreprocessing selanjutnya akan divisualisasikan menggunakan Correlation Heatmap seperti yang terlihat pada Gambar 7. Setiap atribut memiliki nilai korelasi yang berkisar antara 1.00 hingga -1.00. Tahun dan total kejadian per tahun memiliki korelasi yang sangat tinggi sebesar (0.13), bulan dan total kejadian memperlihatkan korelasi sedang (0.13) dan provinsi, numerik penyebab, tanggal memiliki korelasi rendah.

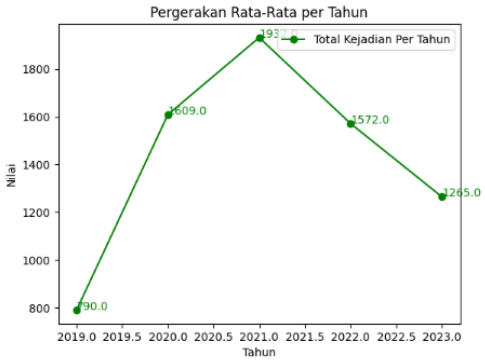


Gambar 7. Korelasi Heatmap

Setelah tahap correlation heatmap, data kemudian di visualisasi kembali menggunakan lineplot untuk menganalisis pergerakan kejadian banjir per tahun, seperti yang terlihat pada gambar 8. Lineplot ini menunjukkan

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 219−228
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5958

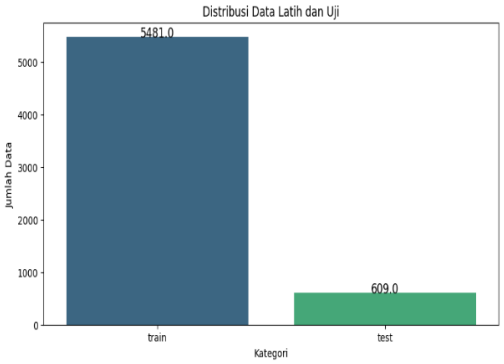
fluktuasi jumlah kejadian banjir setiap tahunnya, dengan nilai terendah terjadi pada tahun 2019, mencapai 790 kejadian. Sementara itu, nilai tertinggi terjadi pada tahun 2021, dengan jumlah 1.932 kejadian. Visualisasi ini membantu dalam mengidentifikasi tren dan pola kejadian banjir dari tahun ke tahun.



Gambar 8. Pergerakan Banjir per tahun

3.5 Split Data

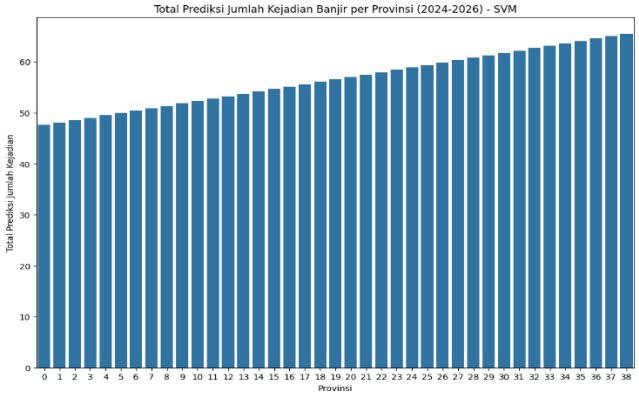
Pada tahap ini, setelah data selesai melalui preprocessing dan processing, langkah selanjutnya adalah memisahkan dataset menjadi data training dan testing. Pemisahan ini dilakukan dengan metode split validation, untuk perbandingan split data menggunakan 90 % Training dan 10 % Test untuk pengujian. Pembagian ini bertujuan untuk memudahkan algoritma seperti random forest dan support vector machine dalam memproses data. Seperti yang terlihat pada gambar 9, dataset training terdiri dari 5481, sedangkan dataset testing terdiri dari 609.



Gambar 9. Split data

3.6 Implementasi Algoritma Support Vector Machine

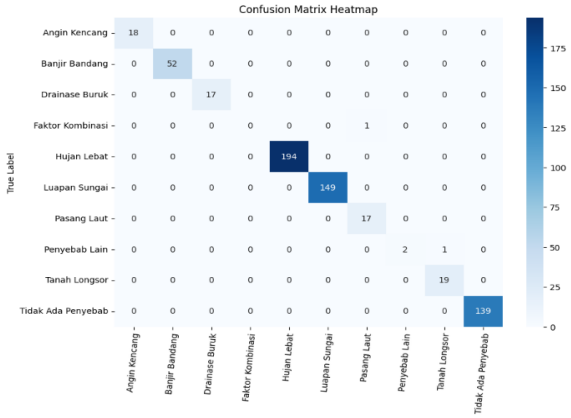
Pada tahap ini, data yang telah melalui beberapa tahap seperti preprocessing, processing dan split data. Pada implementasi algoritma Support Vector Machine untuk menghasilkan prediksi bencana alam di Indonesia dengan visualisasi barplot, pada gambar 10 menunjukkan distribusi total prediksi jumlah kejadian banjir berdasarkan kategori penyebab di setiap provinsi di Indonesia. Dari hasil prediksi tertinggi terdapat sebanyak 4 provinsi pada tahun 2024 – 2026 ialah Sumatera Utara 63 kejadian, Sumatera Selatan 62 kejadian, Sumatera Barat 61 kejadian dan Sulawesi Utara 60 kejadian, dengan penyebab terbesar ialah hujan lebat, banjir bandang dan luapan sungai. Hasil dari penerapan algoritma Support Vector Machine mendapatkan hasil akurasi model sebanyak 99,5%.



Gambar 10. Visualisasi prediksi banjir tahun 2024 – 2026

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 219−228
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5958

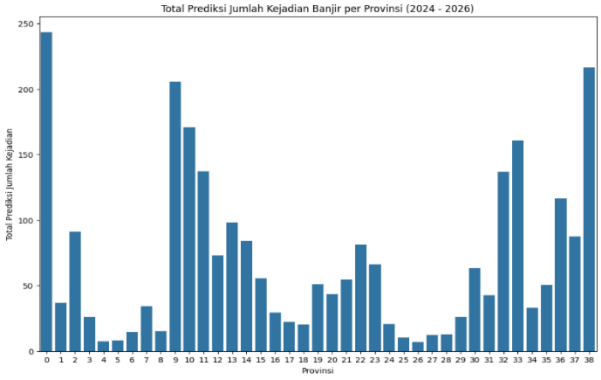
Pada Gambar 11, ditampilkan hasil confusion matrix dari algoritma Support Vector Machine yang menunjukkan hubungan antara label sebenarnya (true label) dan prediksi model dalam beberapa kategori penyebab bencana alam. Prediksi terbesar terlihat pada kategori Hujan Lebat dengan 194 kejadian, sedangkan prediksi terendah terdapat pada kategori Faktor Kombinasi dengan 0 kejadian.



Gambar 11. Confusion matrix SVM

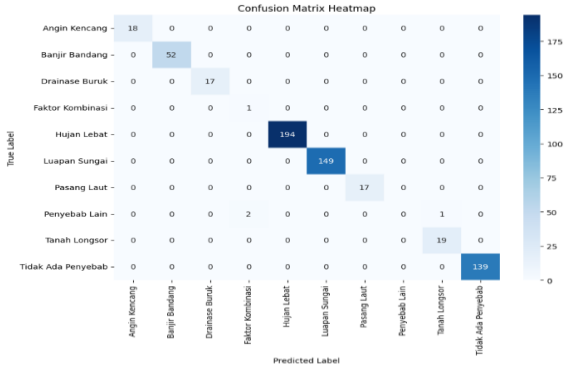
3.7 Implementasi Algoritma Random Forest

Pada tahap ini, data yang telah melalui beberapa tahap seperti preprocessing, processing dan split data. Pada implementasi algoritma Random Forest untuk menghasilkan prediksi bencana alam di Indonesia dengan visualisasi barplot, pada gambar 12 menunjukkan distribusi total prediksi jumlah kejadian banjir berdasarkan kategori penyebab di setiap provinsi di Indonesia. Dari hasil prediksi tertinggi terdapat sebanyak 4 provinsi pada tahun 2024 – 2026 ialah Aceh sebanyak 240 kejadian banjir, Sumatera Utara 215, Jawa Barat sebanyak 210 kejadian banjir dan Jawa Tengah sebanyak 160 kejadian banjir, dengan penyebab terbesar ialah hujan lebat, banjir bandang dan luapan sungai. Hasil dari penerapan algoritma Random Forest mendapatkan hasil akurasi sebesar 99,6%.



Gambar 12. Visualisasi prediksi Random Forest 2024 – 2026

Pada Gambar 13, ditampilkan hasil confusion matrix dari algoritma Support Vector Machine yang menunjukkan hubungan antara label sebenarnya (true label) dan prediksi model dalam beberapa kategori penyebab bencana alam. Prediksi terbesar terlihat pada kategori Hujan Lebat dengan 194 kejadian, sedangkan prediksi terendah terdapat pada kategori penyebab lainnya dengan 0 kejadian.



Gambar 13. Confusion Matrix Random Forest

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 219−228
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5958

3.8 Evaluasi RSME

Hasil akurasi evaluasi RSME menunjukkan bahwa algoritma Random Forest lebih optimal dibandingkan dengan Support Vector Machine. Perbandingan nilai RSME terbaik menunjukkan bahwa Algoritma Random Forest menghasilkan nilai 3.853, sebagaimana ditampilkan pada Gambar 14.

	Model	Rata-rata RMSE
0	Random Forest	3.853083
1	Support Vector Machine	5.959030

Gambar 14. Hasil RSME

4. KESIMPULAN

Pada kesimpulan penelitian ini tentang Analisis Prediksi Banjir di Indonesia Menggunakan Algoritma Support Vector Machine dan Random Forest Studi Kasus BNPB, dengan tahapan penelitian seperti pengumpulan data yang diperoleh di laman web gis.bnpb.co.id yang berisikan 7,168 kejadian pada tahun 2019 – 2023, analisa data yang berisikan missing value, duplikasi data dan presentase data yang dihapus, preprocessing data yang berisikan Cleaning data, Transformasi data, Undersampling dan Ekstraksi fitur, processing data yang berisikan Feature Selection, Pola data pertahun dan Split Data, implementasi algoritma dan evaluasi RMSE. Berdasarkan hasil, penelitian untuk prediksi bencana banjir di provinsi Indonesia, prediksi kejadian banjir yang sering muncul di dua algoritma tersebut ialah Provinsi Sumatera Utara, Aceh, Jawa barat, Jawa Tengah, Sumatera Selatan, Sumatera Barat dan Sulawesi Utara, dengan penyebab yang sering muncul ialah hujan lebat, banjir bandang dan luapan sungai. Adapun hasil akurasi algoritma support vector machine sebesar 99,5% dan algoritma dengan akurasi terbaik menurut Root Square Mean Eror ialah Random Forest sebesar 96,6% dan rata-rata RSME 3.853.

REFERENCES

[1] D. Susanti and T. Wahyuni, “ANALISIS POTENSI BENCANA ALAM TANAH LONGSOR KABUPATEN MAJALENGKA MENGGUNAKAN ALGORITMA NAÏVE BAYES CLASSIFIER,” INFOTECH journal, vol. 9, no. 2, pp. 299–306, Jul. 2023, doi: 10.31949/infotech.v9i2.5645.
[2] M. Althaf Pramasetya Perkasa, “Analisis Probabilitas Bencana Alam dengan Penerapan Data Mining Menggunakan K-Means dan Linier Regression.” 2023
[3] S. H. Hengkelare et al., “MITIGASI RISIKO BENCANA BANJIR DI MANADO,” Jurnal Spasial, vol. 8, no. 2, p. 2021.
[4] D. D. Utomo and F. Y. D. Marta, “Dampak Bencana Alam Terhadap Perekonomian Masyarakat di Kabupaten Tanah Datar,” JURNAL TERAPAN PEMERINTAHAN MINANGKABAU, vol. 2, no. 1, pp. 92–97, Jun. 2022, doi: 10.33701/jtpm.v2i1.2395.
[5] Y. Nuryaman, A. Yudha, and A. Asistyasari, “Analisis Prediksi Bencana Angin Puyuh di Jawabarat menggunakan Algoritma K-NN dan C4.5 Berbasis PSO,” Simposium Nasional Ilmiah dengan tema: (Peningkatan Kualitas Publikasi Ilmiah melalui Hasil Riset dan Pengabdian kepada Masyarakat, pp. 513–514, Nov. 2019.
[6] N. Hidayati, P. T. Pungkasanti, and N. Wakhidah, “Prediksi Bencana Alam di Kota Semarang Menggunakan Algoritma Markov Chains,” Jurnal Sains dan Informatika, vol. 7, no. 1, pp. 107–116, Jul. 2021, doi: 10.34128/jsi.v7i1.283.
[7] A. Salaffudin, N. Nafi’iyah, N. Q. Nawafilah, and U. I. Lamongan, “Algoritma Backpropagation untuk Memprediksi Korban Bencana Alam,” SMATIKA, vol. 9, no. 2087–0256, pp. 77–79, Dec. 2019.
[8] M. Murdiaty, A. Angela, and C. Sylvia, “Pengelompokkan Data Bencana Alam Berdasarkan Wilayah, Waktu, Jumlah Korban dan Kerusakan Fasilitas Dengan Algoritma K-Means,” JURNAL MEDIA INFORMATIKA BUDIDARMA, vol. 4, no. 3, p. 744, Jul. 2020, doi: 10.30865/mib.v4i3.2213.
[9] Y. Ramdhani and A. Mubarok, “Analisis Time Series Prediksi Penutupan Harga Saham Antm.Jk Dengan Algoritma SVM Model Regresi,” JURNAL RESPONSIF, vol. 1, no. 1, Aug. 2019, [Online]. Available: http://ejurnal.univbsi.id/index.php/jti
[10] S. Dwiasnati and Y. Devianto, “Optimasi Prediksi Bencana Banjir menggunakan Algoritma SVM untuk penentuan Daerah Rawan Bencana Banjir,” SISFOTEK, vol. 5, pp. 202–203, Sep. 2021.
[11] Mia, A. F. N. Masruriyah, and A. R. Pratama, “KOMPARASI MODEL DECISION TREE DAN RANDOM FOREST UNTUK MEMPREDIKSI PENYAKIT JANTUNG,” Scientific Student Journal for Information, Technology and Science, vol. 2, no. 2715–2766, p. 126, Jun. 2023.
[12] D. Fitrianah, W. Gunawan, and A. Puspita Sari, “Studi Komparasi Algoritma Klasifikasi C5.0, SVM dan Naive Bayes dengan Studi Kasus Prediksi Banjir Comparative Study of Classification Algorithm between C5.0, SVM and Naive Bayes with Case Study of Flood Prediction,” Feb. 2022.
[13] A. Fitra, “Pengembangan Model Prediksi Masa Studi Sarjana Menggunakan Regresi Linear,” Jun. 2022.
[14] M. Bagas, A. Darmawan, F. Dewanta, and S. Astuti, “Analisis Perbandingan Algoritma Decision Tree, Random Forest, dan Naïve Bayes untuk Prediksi Banjir di Desa Dayeuhkolot Comparative Analysis of Decision Tree, Random Forest, and Naïve Bayes Algorithm for Flood Prediction at Dayeuhkolot Village,” TELKA, vol. 9, no. 1, pp. 52–61, May 2023.
[15] R. Y. Hayuningtyas and R. Sari, “Implementasi Data Mining Dengan Algoritma Multiple Linear Regression Untuk Memprediksi Penyakit Diabetes.,” Jurnal Teknik Komputer AMIK BSI, vol. 8, pp. 40–43, Jan. 2022, doi: 10.31294/jtk.v4i2.

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 219−228
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5958

[16] U. Amelia, J. Indra, and A. F. N. Masruriyah, "IMPLEMENTASI ALGORITMA SUPPORT VECTOR MACHINE(SVM) UNTUK PREDIKSI PENYAKIT STROKEDENGAN ATRIBUT BERPENGARUH," Scientific Student Journal for Information, Technology and Science, no. 2715–2766, Jun. 2024.
[17] H. Badruzzaman, T. Al Mudzakir, and Rahmat, "IMPLEMENTASI ALGORITMA CONVOLUTIONAL NEURAL NETWORKDAN SUPPORT VECTOR MACHINE UNTUK PENDETEKSIAN CANDI JIWA DAN CANDI BLANDONGAN," Scientific Student Journal for Information, Technology and Science, no. 2715–2766, Jun. 2024.
[18] U. Erdiansyah, A. Irmansyah Lubis, and K. Erwansyah, "Komparasi Metode K-Nearest Neighbor dan Random Forest Dalam Prediksi Akurasi Klasifikasi Pengobatan Penyakit Kutil," JURNAL MEDIA INFORMATIKA BUDIDARMA, vol. 6, no. 1, p. 208, Jan. 2022, doi: 10.30865/mib.v6i1.3373.
[19] W. Apriliah et al., "SISTEMASI: Jurnal Sistem Informasi Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," 2021. [Online]. Available: http://sistemasi.ftik.unisi.ac.id
[20] A. A. Nurhalizah, Y. Cahyana, and Rahmat, "Model Prediksi Kekuatan Gempa Dengan Menggunakan Algoritma Linear Regression Dan Support Vector Regression (Studi Kasus BMKG)," no. 2, p. 41, 2024, [Online]. Available: https://www.kaggle.com/datasets/kekavigi/earthquakes-in-ndonesia.