Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024 Page: 1222−1230
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5593

Penerapan Algoritma Random Forest untuk Memprediksi Curah Hujan pada Masa Mendatang di Daerah Berpotensi Banjir

Novie Rahmadani*, Ade Silvia Handayani, Irawan Hadi

Teknik Elektro, Program Studi Teknik Telekomunikasi, Politeknik Negeri Sriwijaya, Palembang, Indonesia
Email: novierahmadani411@gmail.com, ade_silvia@polsri.ac.i, irawanhadi1965@gmail.com
Email Penulis Korespondensi: novierahmadani411@gmail.com
Submitted: **17/07/2024**; Accepted: **27/09/2024**; Published**: 30/09/2024**

Abstrak−Palembang, sebagai salah satu kota terbesar di Indonesia, secara rutin mengalami masalah banjir parah setiap tahunnya. Banjir tidak hanya mengganggu aktivitas sehari-hari penduduk, tetapi juga menyebabkan kerugian ekonomi yang signifikan dan dampak sosial. Untuk mengatasi masalah ini, sangat penting memiliki pemahaman mendalam tentang pola banjir dan beberapa faktor yang mempengaruhinya. Tujuan dari penelitian ini adalah untuk menerapkan teknologi Machine Learning (ML) yang sangat efisien untuk analisis prediksi daerah rawan banjir dimasa mendatang. Integrasi ML dapat membantu dalam mengidentifikasi pola, memprediksi risiko, dan membuat keputusan yang lebih akurat dalam mitigasi banjir. Dalam upaya mencapai tujuan ini, metodologi CRISP-DM (Cross-Industry Standard Process for Data Mining) akan diterapkan untuk memastikan penelitian dilakukan secara sistematis dan komprehensif. Oleh karena itu, penelitian tentang analisis pemetaan daerah rawan banjir di Palembang menggunakan ML sangat penting untuk menyediakan solusi yang cukup efektif dan efisien untuk permasalahan banjir yang telah berlangsung lama. Dengan pendekatan CRISP-DM, diharapkan penelitian ini dapat menghasilkan suatu model prediksi yang akurat dan dapat diandalkan dengan mengintegerasikan algoritma Random Forest sebagai model regresi, serta memberikan manfaat jangka panjang bagi pengelolaan risiko banjir di Palembang dan beberapa kota lain di Indonesia yang mengalami masalah serupa.

Kata Kunci: Banjir; CRISP-DM; Data Mining; Machine Learning; Random Forest

Abstract−Palembang, as one of the largest cities in Indonesia, regularly experiences severe flooding problems every year. Flooding not only disrupts the daily activities of residents, but also causes significant economic losses and social impacts. To solve this problem, it is crucial to have an in-depth understanding of flooding patterns and some of the factors that influence them. The purpose of this research is to apply highly efficient Machine Learning (ML) technology for the prediction analysis of future flood-prone areas. The integration of ML can help in identifying patterns, predicting risks, and making more accurate decisions in flood mitigation. In an effort to achieve this goal, the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology will be applied to ensure the research is conducted systematically and comprehensively. Therefore, research on the analysis of mapping flood-prone areas in Palembang using ML is essential to provide a fairly effective and efficient solution to the long-standing flooding problem. With the CRISP-DM approach, it is expected that this research can produce an accurate and reliable prediction model by integrating the Random Forest algorithm as a regression model, and provide long-term benefits for flood risk management in Palembang and several other cities in Indonesia that experience similar problems.

Keywords: Flood; CRISP-DM; Data Mining; Machine Learning; Random Forest

1. PENDAHULUAN

Beberapa daerah di Indonesia khususnya kota-kota besar yang juga padat pemukiman sering dilanda bencana alam, salah satunya yaitu banjir yang kerap terjadi setiap tahunnya[1]. Dampaknya yang signifikan terhadap kehidupan masyarakat, infrastruktur, dan ekonomi menjadikan prediksi banjir sebagai suatu hal yang sangat krusial[2]. Curah hujan yang tidak menentu dan meningkatnya frekuensi badai tropis dapat menyebabkan banjir yang lebih sering dan lebih parah[3]. Akibatnya, banyak kota di Indonesia yang mengalami kerugian besar baik secara ekonomi maupun sosial. Banjir dapat menyebabkan kerusakan pada rumah, jalan, dan infrastruktur penting lainnya, serta mengganggu aktivitas ekonomi dan menyebabkan penurunan kualitas hidup masyarakat yang terdampak[4]. Oleh sebab itu, diperlukan metode yang efektif untuk memprediksi terjadinya banjir agar dapat dilakukan langkah-langkah mitigasi yang tepat waktu. Prediksi banjir menjadi semakin penting mengingat perubahan iklim yang semakin memperburuk kondisi cuaca ekstrem[5].

Salah satu pendekatan yang dapat digunakan untuk memprediksi banjir adalah dengan menerapkan teknologi Machine Learning (ML)[6]. ML memungkinkan kita untuk menganalisis sejumlah besar data dan mengidentifikasi pola-pola yang mungkin tidak terlihat oleh metode tradisional[7]. Dengan menggunakan data historis tentang kejadian banjir, curah hujan, dan faktor-faktor cuaca lainnya, ML dapat membantu dalam memprediksi kemungkinan terjadinya banjir di masa mendatang[8]. Salah satu algoritma yang dapat digunakan dalam konteks ini adalah Random Forest. Random Forest adalah algoritma pembelajaran ensemble yang menggunakan sejumlah besar pohon keputusan untuk membuat prediksi[9]. Algoritma ini bekerja dengan membuat banyak pohon keputusan selama pelatihan dan mengeluarkan rata-rata prediksi dari masing-masing pohon[10].

Keuntungan dari menggunakan Random Forest adalah kemampuannya untuk menangani dataset yang besar dan kompleks, serta kemampuannya untuk menangani data yang tidak seimbang[11]. Selain itu, Random Forest dapat memberikan estimasi tentang pentingnya variabel, yang dapat membantu kita memahami faktor-faktor utama yang berkontribusi terhadap kejadian banjir. Dalam penerapannya, data historis tentang curah hujan, kelembaban, dan kejadian banjir digunakan untuk melatih model Random Forest. Data tersebut kemudian dibagi menjadi dua set: set pelatihan dan set pengujian. Set pelatihan digunakan untuk membuat model, sedangkan set pengujian digunakan untuk

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024 Page: 1222−1230
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5593

mengevaluasi akurasi model[12]. Setelah model dilatih, kita dapat menggunakan model tersebut untuk memprediksi kemungkinan terjadinya banjir berdasarkan data cuaca terbaru[13].

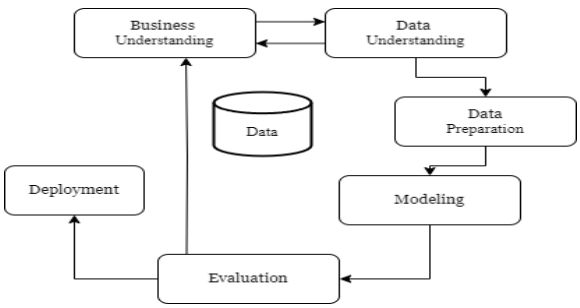
Menurut [14], penggunaan beberapa algoritma ML untuk memprediksi banjir di Jakarta telah dikaji, di mana data historis tentang curah hujan, ketinggian air, dan faktor cuaca lainnya digunakan untuk melatih model. Algoritma yang diuji termasuk Random Forest, Support Vector Machines, dan Neural Networks, dengan hasil menunjukkan bahwa Random Forest memberikan hasil prediksi yang paling akurat. Penelitian [15] menggunakan data spasial dan algoritma Gradient boosting dan Random Forest untuk memprediksi daerah rawan banjir di Surabaya. Penelitian ini berhasil mengidentifikasi daerah yang paling rentan terhadap banjir dengan tingkat akurasi yang tinggi, menggunakan data yang mencakup topografi, penggunaan lahan, dan curah hujan. Penelitian oleh [16] mengaplikasikan metodologi CRISP-DM dan algoritma Random Forest untuk prediksi banjir di Semarang. Data historis banjir, curah hujan, dan data topografi digunakan untuk melatih model, yang menunjukkan bahwa kombinasi CRISP-DM dan Random Forest efektif dalam memprediksi banjir dan memberikan informasi yang bermanfaat untuk mitigasi bencana. Peneliti [17]menggunakan Neural Networks untuk memprediksi banjir di Medan dengan menggunakan data cuaca seperti curah hujan, suhu, dan kelembaban. Penelitian ini menunjukkan bahwa Neural Networks dapat mengidentifikasi pola yang kompleks dalam data cuaca dan memberikan prediksi yang akurat mengenai kemungkinan terjadinya banjir. Selain itu, penelitian oleh [18] mengkaji prediksi tahap banjir di Sungai Parma di Italia menggunakan Support Vector Regression yang menunjukkan bahwa model LSTM adalah yang paling akurat dalam memprediksi nilai puncak banjir.

Selain itu, penting untuk melibatkan berbagai pihak dalam upaya mitigasi banjir, termasuk pemerintah, komunitas lokal, dan sektor swasta. Kolaborasi antara berbagai pihak ini dapat membantu dalam pengumpulan data yang lebih lengkap dan akurat, serta dalam penerapan langkah-langkah mitigasi yang lebih efektif. Misalnya, pemerintah dapat menyediakan data cuaca dan topografi, sementara komunitas lokal dapat memberikan informasi tentang kondisi lingkungan setempat dan sejarah kejadian banjir. Sektor swasta, terutama perusahaan teknologi, dapat memberikan dukungan dalam hal penyediaan perangkat keras dan perangkat lunak yang diperlukan untuk analisis data dan pembuatan model prediksi. Edukasi dan peningkatan kesadaran masyarakat tentang pentingnya mitigasi banjir juga sangat penting. Masyarakat perlu diberikan informasi tentang langkah-langkah yang dapat diambil untuk mengurangi risiko banjir, seperti membangun rumah di lokasi yang lebih tinggi, menggunakan bahan bangunan yang tahan air, dan memastikan sistem drainase yang baik. Dengan demikian, masyarakat dapat lebih siap dalam menghadapi banjir dan mengurangi dampak yang ditimbulkannya. Dengan penerapan teknologi ML dan kolaborasi yang baik antara berbagai pihak, diharapkan prediksi banjir dapat dilakukan dengan lebih akurat dan efektif. Hal ini akan membantu dalam upaya mitigasi banjir dan mengurangi risiko serta dampak yang diakibatkan oleh bencana ini. Selain itu, pendekatan ini juga dapat diterapkan di kota-kota lain di Indonesia yang menghadapi masalah serupa, sehingga manfaatnya dapat dirasakan secara luas.

Dalam jangka panjang, penelitian ini diharapkan dapat memberikan kontribusi yang signifikan dalam pengelolaan risiko banjir di Indonesia. Dengan model prediksi yang akurat dan langkah-langkah mitigasi yang efektif, kita dapat mengurangi kerugian ekonomi dan sosial yang diakibatkan oleh banjir, serta meningkatkan kualitas hidup masyarakat yang terdampak. Selain itu, pendekatan ini juga dapat membantu dalam perencanaan pembangunan yang lebih berkelanjutan, dengan mempertimbangkan risiko banjir dan faktor-faktor lingkungan lainnya. Pada akhirnya, upaya mitigasi banjir yang efektif akan membantu menciptakan masyarakat yang lebih tangguh dan siap menghadapi bencana alam, serta mendukung pembangunan yang lebih berkelanjutan dan inklusif di Indonesia.

2. METODOLOGI PENELITIAN

Penelitian ini mengimplementasikan metodologi data mining CRISP-DM sebagai kerangka kerja umum untuk mengatasi masalah dalam penelitian. Metodologi ini memberikan kerangka kerja yang terstruktur dan terukur untuk menjalankan proyek data mining dari awal hingga akhir. Dalam metodologi ini diimplementasikan algoritma Random Forest untuk memodelkan regresi. CRISP-DM terdiri dari enam tahapan utama yang mencakup seluruh proses data mining: Business Understanding (Pemahaman Bisnis), Data Understanding (Pemahaman Data), Data Preparation (Persiapan Data), Modeling (Pemodelan), Evaluation (Evaluasi), dan Deployment (Penerapan). Dengan mengikuti metodologi CRISP-DM, penelitian ini diharapkan dapat mengatasi masalah dengan pendekatan yang sistematis dan menghasilkan solusi data mining yang efektif serta dapat diandalkan untuk mendukung tujuan bisnis dan penelitian.



Gambar 1. Metodologi CRISP-DM

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024 Page: 1222−1230
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5593

2.1 Business Understanding

Tahap ini fokus pada pemahaman masalah bisnis dan tujuan yang ingin dicapai melalui proyek data mining[19]. Langkah-langkah yang dilakukan termasuk mengidentifikasi tujuan bisnis, menetapkan tujuan data mining yang spesifik, serta mengevaluasi sumber daya yang tersedia dan kendala yang mungkin muncul.

2.2 Data Understanding

Tahap ini merupakan tahap penting dalam proses data mining, sangat penting untuk meletakkan fondasi yang kuat sebelum pemodelan atau analisis dimulai[20]. Fase ini adalah didedikasikan untuk memahami seluk-beluk dan kualitas data secara komprehensif dataset yang diperuntukkan untuk proyek tersebut. Awalnya, ini melibatkan pemeriksaan menyeluruh untuk mengungkap pola yang melekat, mendeteksi anomali seperti nilai yang hilang dan outlier, dan menilai dampak potensial mereka pada analisis selanjutnya. Di luar pemeriksaan awal ini, proses ini menggali lebih dalam setiap atribut yang ada yang ada dalam kumpulan data[21]. Visualisasi dan statistik deskriptif memainkan peran kunci selama fase ini fase ini, membantu dalam eksplorasi distribusi data, mengidentifikasi korelasi antar variabel, dan mengungkap hubungan laten yang mungkin mempengaruhi model prediktif[22]. Pemeriksaan yang cermat seperti itu memastikan bahwa data yang yang digunakan tidak hanya dapat diandalkan tetapi juga relevan untuk membangun model prediksi yang akurat dan model prediksi yang kuat. Dengan mendapatkan pemahaman yang mendalam tentang dataset nuansa dan karakteristik set data, ilmuwan data dapat secara efektif mengurangi bias, memvalidasi asumsi, dan membuat keputusan yang tepat mengenai langkah-langkah prapemrosesan dan strategi rekayasa fitur. Pengetahuan mendasar ini bertindak sebagai landasan, mengarahkan tahap selanjutnya dari proses persiapan data, pemodelan dan evaluasi untuk mencapai wawasan yang bermakna dan hasil yang dapat ditindaklanjuti dalam dalam ranah pengambilan keputusan berbasis data.

2.3 Data Preparation

Langkah penting dalam siklus data mining adalah persiapan data, yang meliputi mempersiapkan data mentah ke dalam bentuk yang sesuai untuk analisis lebih lanjut. Langkah-langkah yang dilakukan meliputi pembersihan data dari nilai yang hilang atau outlier yang dapat mengganggu analisis, transformasi data untuk memenuhi asumsi model, dan penggabungan atau pengelompokan data jika diperlukan[23]. Selain itu, tahap ini juga mencakup normalisasi atau standarisasi data untuk memastikan bahwa variabel-variabel memiliki skala yang seragam. Persiapan data yang cermat dan teliti sangat penting untuk menjamin bahwa model yang dibangun dapat memberikan hasil yang akurat dan dapat dipercaya serta didasarkan pada data yang berkualitas tinggi. Dalam tahap persiapan data, beberapa langkah penting dilakukan untuk memastikan kualitas data yang akan digunakan dalam analisis. Salah satu langkah yang dilakukan adalah pembersihan data, yaitu penanganan nilai yang hilang. Nilai yang hilang ditangani dengan mengisi nilai yang hilang menggunakan metode yang sesuai atau menghapus data yang hilang jika tidak terlalu signifikan[24]. Selain itu, tipe data yang tidak sesuai juga diperbaiki dengan cara mengubah tipe data agar sesuai dengan kebutuhan analisis. Dan jika memungkinkan, kolom-kolom yang tidak diperlukan dalam analisis akan dihapus untuk menyederhanakan dataset dan meningkatkan efisiensi proses pemodelan.

2.4 Modeling

Salah satu fase penting lainnya dalam proses data mining adalah proses pemodelan di mana model atau metode analisis data diterapkan pada dataset yang telah disiapkan sebelumnya. Langkah-langkahnya meliputi pemilihan model yang paling sesuai dengan tujuan analisis dan karakteristik data, serta membagi dataset menjadi data untuk pengujian dan pelatihan untuk evaluasi model[25]. Selanjutnya, model yang dipilih dinilai kinerjanya setelah dilatih pada set data awal. menggunakan metrik evaluasi yang sesuai. Tahap ini juga melibatkan tuning parameter model dan teknik validasi silang untuk meningkatkan kinerja model. Hasil dari tahap ini adalah model prediktif yang dapat digunakan untuk membuat prediksi atau mengidentifikasi pola yang berguna dalam data. Dalam penelitian ini, library scikit-learn digunakan dengan algoritma Random Forest untuk menganalisis data dan membangun model prediktif[26]. Scikit-learn menyediakan implementasi yang sangat baik untuk Random Forest, lengkap dengan berbagai alat untuk pemodelan data dan evaluasi model. Selain itu, Random Forest memungkinkan analisis mendalam tentang kontribusi setiap fitur terhadap variabel target.

2.5 Evaluation

Tahap Evaluation dalam konteks regresi atau prediksi sangat penting untuk mengukur kualitas dan akurasi model yang telah dibangun dalam memprediksi nilai kontinu[27]. Dalam tahap evaluasi regresi atau prediksi, selain metrik seperti Mean Squared Error (MSE), Mean Absolute Error (MAE), dan koefisien determinasi (R^2), metrik lainnya yang sering digunakan adalah Root Mean Squared Error (RMSE) dan Mean Absolute Percentage Error (MAPE). RMSE adalah akar dari MSE dan memberikan gambaran tentang seberapa besar kesalahan rata-rata dari prediksi model dalam unit yang sama dengan variabel target. Sedangkan MAPE mengukur persentase rata-rata kesalahan dari prediksi model terhadap nilai aktual, memberikan gambaran tentang tingkat kesalahan relatif dari model. Evaluasi model juga dapat melibatkan teknik validasi silang untuk memastikan bahwa model memiliki kinerja yang konsisten dan dapat diandalkan saat diimplementasikan pada data baru yang belum pernah dilihat sebelumnya. Hasil dari tahap evaluasi

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024 Page: 1222−1230
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5593

ini memberikan wawasan yang berharga tentang kecocokan model dengan tujuan bisnis yang ditetapkan serta membantu dalam pengambilan keputusan terkait penyesuaian atau penyempurnaan model jika diperlukan.

2.6 Deployment

Tahap Deployment merupakan langkah kunci dalam proses data mining di mana model yang telah dikembangkan dan dievaluasi siap untuk diimplementasikan dalam lingkungan produksi. Selama tahap ini, model yang telah disetujui dipersiapkan untuk digunakan secara praktis dalam pengambilan keputusan sehari-hari. Ini melibatkan integrasi model ke dalam sistem atau aplikasi yang relevan, serta pengujian ulang untuk memastikan bahwa model dapat beroperasi dengan baik dalam situasi dunia nyata. Selain itu, proses deployment juga mencakup pelatihan pengguna yang akan menggunakan model, dokumentasi yang tepat, dan proses pemantauan dan pemeliharaan model untuk memastikan kinerjanya.

3. HASIL DAN PEMBAHASAN

Banjir yang sering terjadi di Kota Palembang berdampak besar pada kehidupan warga dan infrastruktur kota. Terutama saat musim hujan, dengan curah hujan yang tinggi di daerah hulu atau sekitar Palembang, Sungai Musi dapat meluap dan menyebabkan kenaikan permukaan air secara tiba-tiba. Kondisi ini sering kali menyebabkan banjir di berbagai wilayah di Palembang. Oleh karena itu, kebutuhan akan prediksi banjir di Palembang menjadi sangat penting untuk mengantisipasi dan mengurangi dampak dari fenomena banjir yang sering terjadi. Dengan menggunakan metode prediksi seperti CRISP-DM dan algoritma Random Forest, kita dapat mengembangkan model yang mampu memprediksi tingkat risiko banjir berdasarkan faktor-faktor seperti curah hujan, tinggi muka air sungai, drainase, dan topografi. Prediksi yang akurat dapat memberikan informasi yang berharga bagi pemerintah dan pemangku kepentingan terkait untuk merencanakan langkah-langkah mitigasi yang tepat waktu dan efektif.

3.1 Business Understanding

Banjir yang sering terjadi di Kota Palembang berdampak besar pada kehidupan warga dan infrastruktur kota. Terutama saat musim hujan, dengan curah hujan yang tinggi di daerah hulu atau sekitar Palembang, Sungai Musi dapat meluap dan menyebabkan kenaikan permukaan air secara tiba-tiba. Kondisi ini sering kali menyebabkan banjir di berbagai wilayah di Palembang. Oleh karena itu, kebutuhan akan prediksi banjir di Palembang menjadi sangat penting untuk mengantisipasi dan mengurangi dampak dari fenomena banjir yang sering terjadi. Dengan menggunakan metode prediksi seperti CRISP-DM dan algoritma Random Forest, kita dapat mengembangkan model yang mampu memprediksi tingkat risiko banjir berdasarkan faktor-faktor seperti curah hujan, tinggi muka air sungai, drainase, dan topografi. Prediksi yang akurat dapat memberikan informasi yang berharga bagi pemerintah dan pemangku kepentingan terkait untuk merencanakan langkah-langkah mitigasi yang tepat waktu dan efektif.

3.2 Data Understanding

Dataset ini diperoleh dari situs web database BMKG yaitu dataonline.bmkg.go.id, khususnya dari stasiun klimatologi Sumatera Selatan di kota Palembang. Informasi yang dikumpulkan meliputi berbagai data klimatologi yang dicatat secara periodik oleh stasiun tersebut. Data yang digunakan dalam penelitian ini mencakup berbagai jenis informasi yang relevan untuk memahami dan memprediksi potensi banjir di wilayah studi.

Tabel 1. Dataset penelitian

Tanggal	Tn	Tx	Tavg	RH avg	RR	SS	ff_x	ddd_x	ff_avg	ddd_car
01-01-2024	25,8	31,8	29,4	77	5	4,5	3	277	1	W
02-01-2024	25,4	33	30,4	74	5	9,3	2	297	1	W
03-01-2024	25,8	31,4	28,9	81	5	9,8	3	307	2	NW
04-01-2024	25,2	29,8	27,5	87	0	8,2	3	278	1	C
05-01-2024	25,6	34,2	30,2	73	7,5	7	4	118	2	W
06-01-2024	24,8	30,5	27,8	85	9,7	8,7	2	279	0	C
07-01-2024	24,2	32,2	29,6	76	7	0	3	250	1	W
08-01-2024	25,6	32,7	29,9	76	0	9,3	3	213	2	W
09-01-2024	26,2	31,5	29,6	81	0	10,8	3	262	0	C
10-01-2024	24,4	30,7	28,6	82	22	10,8	3	270	1	W
...	...	...	...	...	...	...	...	...	...	...
21-06-2024	25,6	31,6	29,8	76	0	0	2	152	1	C
22-06-2024	26,3	34,2	31,2	70	5	0	2	108	1	SE
23-06-2024	25,7	34,8	31,5	64	5	0	5	139	1	SE
24-06-2024	26,5	33,6	28,3	81	5	0	3	100	0	C
25-06-2024	25,2	31,4	29	81	20,5	0	2	112	0	C
26-06-2024	25,5	35,4	31,6	71	0	0	1	130	1	C
27-06-2024	26,2	34,6	30,3	77	0	0	4	81	1	C

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024 Page: 1222−1230
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5593

Tanggal	Tn	Tx	Tavg	RH avg	RR	SS	ff_x	ddd_x	ff_avg	ddd_car
28-06-2024	25,8	34,6	31,1	74	29,4	0	4	129	1	C
29-06-2024	24,4	33	29,8	76	0	0	2	103	0	C
30-06-2024	25,4	33,2	29,6	77	5	0	3	145	1	C

Dari tabel 1 berisi informasi mengenai dataset yang digunakan pada penelitian ini, berikut merupakan penjelasan setiap variabel.

a. Tanggal
Tanggal mencatat waktu spesifik ketika pengamatan dilakukan. Hal ini penting untuk memantau perubahan kondisi lingkungan dari waktu ke waktu, yang dapat memberikan wawasan tentang pola dan tren yang berkaitan dengan risiko banjir.

b. Temperatur Minimum (Tn)
Temperatur minimum (Tn) adalah nilai terendah dari suhu udara yang tercatat dalam rentang waktu tertentu, biasanya dalam periode 24 jam. Pengukuran ini penting dalam meteorologi dan klimatologi untuk memahami variasi suhu harian dan musiman di suatu lokasi. Tn biasanya terjadi pada saat pagi hari menjelang matahari terbit, di mana suhu udara biasanya mencapai titik terendah sebelum terkena paparan sinar matahari. Data Tn digunakan untuk memprediksi kejadian cuaca ekstrem seperti embun beku atau pembekuan tanah, serta untuk mengukur suhu minimum rata-rata dalam periode yang lebih panjang seperti bulan atau tahun.

c. Temperatur Maximum (Tx)
Temperatur maksimum (Tx) adalah nilai tertinggi dari suhu udara yang tercatat dalam suatu periode waktu tertentu, biasanya dalam rentang 24 jam. Pengukuran ini penting dalam meteorologi untuk memahami variasi suhu harian dan musiman di suatu lokasi. Tx sering terjadi pada siang hari saat matahari berada di puncaknya, di mana paparan sinar matahari membuat suhu udara mencapai titik tertinggi sebelum suhu mulai menurun menjelang malam hari. Data Tx digunakan untuk memprediksi kejadian cuaca panas ekstrem, mengukur suhu maksimum rata-rata dalam periode yang lebih panjang seperti bulan atau tahun, serta memahami tren perubahan iklim di suatu daerah.

d. Temperatur Rata-Rata (Tavg)
Temperatur rata-rata (Tavg) adalah nilai tengah dari suhu udara yang dihitung berdasarkan periode waktu tertentu, seperti harian, bulanan, atau tahunan. Pengukuran Tavg diperoleh dengan menjumlahkan semua nilai suhu dalam rentang waktu tersebut, kemudian dibagi dengan jumlah pengamatan. Misalnya, untuk menghitung Tavg harian, semua nilai suhu harian dijumlahkan dan dibagi dengan jumlah hari dalam periode tersebut.
Tavg penting dalam meteorologi dan klimatologi karena memberikan gambaran umum tentang suhu yang dapat diharapkan dalam periode tertentu di suatu lokasi. Data Tavg membantu dalam membandingkan suhu rata-rata antara lokasi yang berbeda, memantau perubahan suhu jangka panjang akibat perubahan iklim, serta memprediksi tren cuaca atau musim di masa depan.

e. Kelembapan Rata-Rata (RH avg)
Kelembapan rata-rata (RH avg) mengacu pada nilai tengah dari kelembapan udara yang diukur dalam suatu periode waktu tertentu, seperti harian, bulanan, atau tahunan. Kelembapan udara diukur sebagai persentase dari kadar uap air yang terkandung dalam udara relatif terhadap jumlah maksimal yang bisa diakomodasi udara pada suhu yang sama.
Untuk menghitung RH avg, semua nilai kelembapan yang diukur selama periode waktu tersebut dijumlahkan, kemudian dibagi dengan jumlah pengamatan. Data RH avg penting dalam meteorologi karena kelembapan udara mempengaruhi kenyamanan termal manusia dan berbagai proses cuaca. RH avg juga digunakan untuk memahami pola kelembapan dalam berbagai musim, memprediksi kemungkinan terjadinya hujan atau embun, serta mengevaluasi dampak kelembapan terhadap pertanian, kesehatan manusia, dan lingkungan secara umum.

f. Curah Hujan (RR)
Curah hujan (RR) dalam dataset BMKG merujuk kepada data mengenai jumlah air hujan yang tercatat jatuh dalam periode waktu tertentu di berbagai lokasi di Indonesia. Data ini biasanya diukur dalam milimeter (mm) per satuan waktu, seperti per jam, per hari, atau per bulan, tergantung pada jenis data yang tersedia. Data curah hujan dari BMKG sangat penting untuk memantau pola hujan dalam skala harian, musiman, dan tahunan di seluruh Indonesia. Informasi ini digunakan untuk berbagai keperluan, termasuk pemantauan potensi banjir.

g. Lamanya Penyinaran Matahari (SS)
Lamanya Penyinaran Matahari (SS) mengacu pada jumlah waktu dalam sehari ketika sinar matahari mencapai permukaan bumi atau lokasi tertentu. Biasanya diukur dalam jam atau menit, SS adalah parameter penting dalam meteorologi karena mempengaruhi kondisi cuaca dan iklim di suatu tempat. Data SS digunakan untuk memperkirakan jumlah energi matahari yang diterima di suatu wilayah dalam rentang waktu tertentu.

h. Kecepatan Angin Maximum (ff_x)
Kecepatan angin maksimum (ff_x) merujuk kepada nilai tertinggi dari kecepatan angin yang tercatat dalam suatu periode waktu tertentu, seperti per jam atau per hari. Pengukuran ini penting dalam meteorologi untuk memahami intensitas dan potensi bahaya angin kencang di suatu lokasi

i. Arah Angin Saat Kecepatan Maximum (ddd_x)
Arah angin saat kecepatan maksimum (ddd_x) merujuk kepada arah atau orientasi angin pada saat kecepatan angin mencapai nilai tertinggi dalam suatu periode waktu tertentu, seperti per jam atau per hari. Pengukuran ini

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024 Page: 1222−1230
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5593

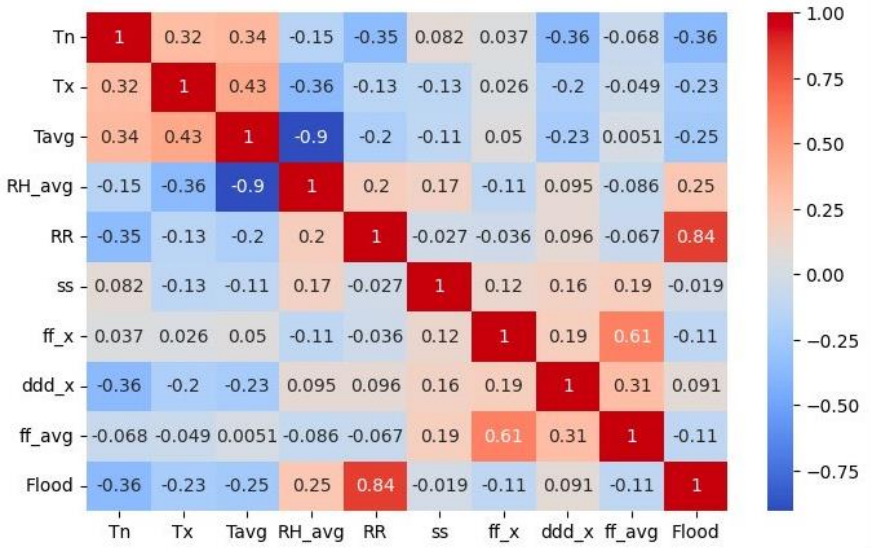
memberikan informasi tambahan yang penting dalam analisis meteorologi karena membantu memahami pola dan sifat dari angin kencang yang terjadi.

j. Kecepatan Angin Rata-Rata (ff_avg)
Kecepatan angin rata-rata (ff_avg) adalah nilai tengah dari kecepatan angin yang diukur dalam suatu periode waktu tertentu, seperti per jam, per hari, atau per bulan. Pengukuran ini penting dalam meteorologi untuk memberikan gambaran umum tentang kecepatan angin yang dapat diharapkan di suatu lokasi dalam jangka waktu tertentu.

k. Arah Angin Terbanyak (ddd_car)
Arah angin terbanyak (ddd_car) mengacu pada arah dominan atau arah utama dari mana angin bertiup dalam suatu periode waktu tertentu, seperti per hari atau per bulan. Data ini memberikan informasi tentang pola dominan dari arah angin yang umumnya terjadi di lokasi tersebut selama periode pengamatan.

3.3 Data Preparation

Pada tahap ini, analisis korelasi antar variabel digunakan untuk memahami hubungan antar variabel dalam kumpulan data. Matriks korelasi membantu dalam mengeksplorasi dan mengekstraksi informasi tentang pola hubungan antar variabel, termasuk seberapa dekat dan arah hubungan tersebut. Dengan memahami korelasi antar variabel, peneliti atau analis data dapat mengidentifikasi pola potensial atau asosiasi yang menarik, serta memastikan bahwa data dipersiapkan dengan baik untuk tahap analisis lebih lanjut.



Gambar 2. Matriks Korelasi

Dari **Tabel 1** dataset penelitian, didapatkan hasil korelasi seperti gambar 2, bahwa bahwa korelasi terbesar terjadi antara variabel Flood dan RR, dengan nilai korelasi sebesar 0.84 atau 84%. Hal ini menunjukkan adanya hubungan yang sangat kuat antara kejadian banjir dengan variabel RR, yang dapat mewakili faktor-faktor seperti curah hujan atau debit air. Korelasi yang tinggi ini mengindikasikan bahwa peningkatan variabel RR secara signifikan berhubungan dengan peningkatan frekuensi atau intensitas banjir. Sehingga saya dapat menjadikan RR sebagai "target" untuk prakiraan dalam tahap pemodelan.

3.4 Modeling

Pada penelitian ini, penjelasan mengenai penggunaan parameter test_size=0.3 dan random_state=42 dalam konteks pembagian data untuk pemodelan atau evaluasi model pada machine learning, khususnya pada pemodelan Random Forest menjadi sangat penting. Parameter test_size=0.3 menentukan proporsi data yang dialokasikan sebagai data uji. Pada contoh ini, nilai 0.3 berarti 30% data akan digunakan sebagai data uji, sedangkan 70% sisanya akan menjadi data latih. Proporsi ini dapat disesuaikan sesuai kebutuhan untuk mengoptimalkan distribusi data antara model pelatihan dan pengujian.

Sementara itu, random_state = 42 merupakan parameter yang digunakan untuk mengontrol pengacakan yang terjadi pada saat pembagian data. Dengan menentukan nilai random_state, maka proses data sharing akan memberikan hasil yang konsisten setiap kali kode dieksekusi dengan nilai yang sama. Hal ini penting untuk memastikan bahwa data sharing tidak berubah secara acak, sehingga hasil evaluasi atau pengujian model dapat direproduksi secara konsisten. Penggunaan kedua parameter ini memungkinkan kita untuk melakukan pembagian data yang terkontrol dan dapat diulang, yang merupakan praktik umum dalam pengembangan model pembelajaran mesin. Hal ini membantu memastikan keandalan dan konsistensi hasil evaluasi model, serta memfasilitasi proses pengembangan model yang lebih terstruktur dan efektif dalam penelitian ini.

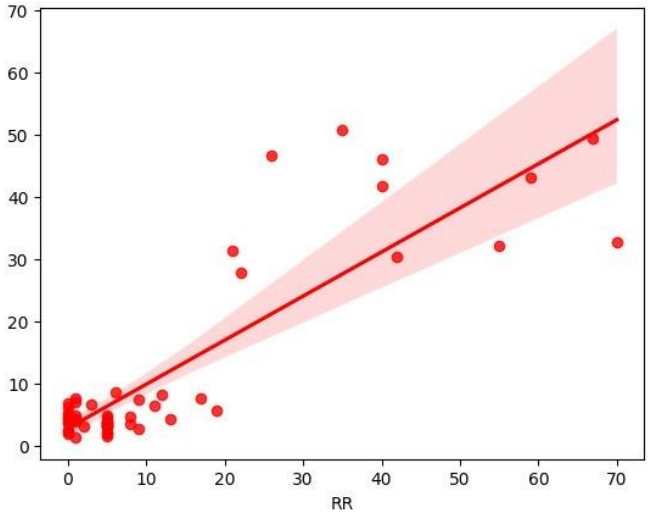
Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024 Page: 1222−1230
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5593

```
[574] y_pred_RF = rf_regressor.predict(x1_test)
      y_pred_RF

array([ 3.58, 41.65,  1.99,  5.62,  8.19,  7.59,  2.07,  4.89,  3.54,
       31.37,  6.24,  3.08,  3.19,  2.74,  3.94, 46.59,  3.12,  2.46,
       27.76,  5.  ,  3.77, 43.1 ,  4.42,  3.48,  3.79,  4.31,  3.84,
        7.1 ,  5.26,  7.72,  7.44,  6.73,  1.47,  4.96, 32.74,  4.55,
       49.28,  2.18, 50.69,  3.96, 32.09,  3.87,  4.3 ,  6.91,  8.64,
        3.21, 30.38,  6.45,  1.63,  4.64, 46.1 ,  4.36,  2.66,  4.49])
```

Gambar 3. Hasil prediksi dalam bentuk array

Gambar 3 menampilkan hasil prediksi Random Forest dalam bentuk array, yang merupakan representasi visual dari proses prediksi curah hujan pada **Tabel 1.** Dataset penelitian variabel RR menggunakan algoritma Random Forest. Setiap elemen dalam array merepresentasikan nilai prediksi curah hujan untuk periode waktu tertentu, yang didasarkan pada set data yang diimplementasikan.



Gambar 4. Visualisasi Prediksi

Gambar 4 menampilkan visualisasi prediksi algoritma Random Forest dalam bentuk plot. Plot ini menunjukkan bagaimana model Random Forest memprediksi nilai curah hujan di Kota Palembang yang sudah dimodelkan berdasarkan dataset pada **Tabel 1.** variabel RR. Pada plot tersebut, terdapat dua garis utama: garis prediksi (dari model Random Forest) dan garis data aktual (data curah hujan yang sebenarnya). Dengan membandingkan kedua garis ini, peneliti atau pemangku kepentingan dapat mengevaluasi seberapa baik model Random Forest memetakan pola curah hujan dari waktu ke waktu. Mereka dapat melihat apakah model mengikuti tren yang sama dengan data aktual dan seberapa dekat prediksi model dengan nilai aktual pada setiap titik waktu.

Lebih lanjut, melalui plot ini, peneliti juga dapat mengidentifikasi pola atau fluktuasi yang mungkin terjadi di masa mendatang. Misalnya, jika ada periode waktu tertentu di mana model tampak kurang akurat, hal ini dapat menjadi perhatian khusus untuk analisis lebih lanjut. Selain itu, peneliti dapat mencari tahu apakah ada pola musiman yang konsisten, seperti peningkatan curah hujan pada musim tertentu, yang dapat diidentifikasi dan diprediksi oleh model. Peneliti juga bisa menggunakan informasi dari plot ini untuk melakukan penyesuaian atau penyempurnaan lebih lanjut pada model Random Forest. Jika ditemukan adanya pola kesalahan atau deviasi yang konsisten, model dapat dioptimalkan dengan menyesuaikan parameter atau dengan menggabungkan data tambahan. Hal ini bertujuan untuk meningkatkan akurasi dan reliabilitas prediksi model di masa mendatang.

3.5 Evaluation

Pustaka Scikit-learn berisi metrik evaluasi yang digunakan dalam analisis regresi untuk mengukur kinerja model prediksi yang terdiri dari Mean Absolute Error (MAE), Mean Squared Error (MSE), Nilai R^2 (Koefisien Determinasi), Root Mean Squared Error (RMSE), dan Mean Absolute Percentage Error (MAPE).
Mean Absolute Error (MAE): Perbedaan absolut rata-rata antara angka yang diharapkan dan angka aktual adalah 5,98907, yang merupakan nilai MAE. Terlepas dari arah kesalahan, MAE memberikan indikator yang jelas mengenai akurasi prediksi. Dalam hal ini, proyeksi sering kali berbeda 5,99 unit dari hasil aktual. Kesalahan Kuadrat Rata-rata (Mean Squared Error/MSE): Dengan merata-ratakan perbedaan kuadrat antara angka yang diharapkan dan angka aktual, nilai MSE sebesar 80,25519 ditentukan. Karena MSE mengkuadratkan setiap perbedaan, MSE menyoroti

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024 Page: 1222−1230
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5593

kesalahan yang lebih besar dan oleh karena itu lebih rentan terhadap outlier atau penyimpangan yang signifikan. Rata-rata kesalahan kuadrat dalam kasus ini adalah 80,26, yang menunjukkan seberapa besar varians kesalahan prakiraan. Root Mean Squared Error (RMSE): RMSE memberikan besaran rata-rata kesalahan dalam unit yang sama dengan data asli. Ini dihitung dengan mengambil akar kuadrat dari kesalahan kuadrat rata-rata (MSE) dalam hal ini, 8,95852. Karena berada pada skala yang sama dengan data, RMSE lebih mudah dibaca daripada MSE. Dalam kasus ini, RMSE menunjukkan bahwa ada penyimpangan rata-rata sebesar 8,96 unit antara proyeksi dan nilai aktual. R-kuadrat (R^2): Persentase varians dalam variabel dependen (nilai aktual) yang dapat diantisipasi berdasarkan variabel independen (nilai prediksi) ditunjukkan oleh nilai R^2 sebesar 0,748416. Pada skala 0 hingga 1, 1 menunjukkan kecocokan yang ideal.

3.6 Deployment

Setelah tahap evaluasi, di mana hasil dari sebuah model dinilai dan dianalisis secara rinci untuk memastikan akurasi dan keandalannya, implementasi dari keseluruhan model yang telah dibangun dan disempurnakan dengan cermat dilakukan. Selain itu, penyesuaian komprehensif juga dilakukan pada model untuk menyempurnakan parameter dan meningkatkan kinerjanya, memastikan bahwa model tersebut memberikan hasil yang sesuai dengan target dan tujuan awal yang telah ditetapkan pada tahap perencanaan tahap CRISP-DM ini.

4. KESIMPULAN

Pada penelitian ini, daerah rawan banjir diprediksi menggunakan metode CRISP-DM dengan algoritma Random Forest. Metode Cross-Industry Standard Process for Data Mining dipilih karena merupakan metodologi terstruktur yang telah terbukti efektif dalam berbagai proyek data mining. Metode ini memberikan panduan langkah demi langkah mulai dari pemahaman bisnis dan data, persiapan data, pemodelan, evaluasi, hingga implementasi. Dengan CRISP-DM, proses prediksi daerah rawan banjir menjadi lebih sistematis dan terarah, sehingga memastikan bahwa setiap langkah dilakukan dengan cermat untuk mencapai hasil yang optimal. Dataset yang digunakan terdiri dari sebelas atribut yaitu Tanggal, Temperatur Minimum (Tn), Temperatur Maksimum (Tx), Temperatur Rata-Rata (Tavg), Kelembapan Rata-Rata (RH avg), Curah Hujan (RR), Lamanya Penyinaran Matahari (SS), Kecepatan Angin Maksimum (ff_x), Arah Angin Saat Kecepatan Maksimum (ddd_x), Kecepatan Angin Rata-Rata (ff_avg), dan Arah Angin Terbanyak (ddd_car). Pustaka scikit-learn digunakan dalam penelitian ini untuk mengimplementasikan algoritma Random Forest. Scikit-learn merupakan library machine learning yang populer di Python, yang menyediakan berbagai alat sederhana dan efisien untuk analisis data dan pemodelan prediktif. Scikit-learn dipilih karena kemudahan penggunaan, dokumentasi yang lengkap, dan performa yang handal dalam memproses data dan membangun model machine learning. Random Forest bekerja dengan menentukan hubungan linier antara variabel independen (seperti tanggal, lokasi, garis bujur, garis lintang, dll.) dan variabel dependen (RR - curah hujan). Model ini mencoba untuk menemukan garis terbaik yang meminimalkan kesalahan prediksi antara nilai aktual dan prediksi. Dalam hal ini, model Random Forest digunakan untuk memprediksi nilai RR berdasarkan atribut lain dalam dataset. Hasil pengujian menunjukkan bahwa algoritma ini bekerja dengan cukup baik dengan nilai R^2 sebesar 74,84%. Prediksi yang dihasilkan memungkinkan untuk menentukan target curah hujan di setiap lokasi di masa depan, sehingga dapat meningkatkan kesiapsiagaan terhadap potensi banjir.

REFERENCES

[1] D. Ayu, H. Sari, J. Rahayu, and B. S. Pujantiyo, “Kajian Kesesuaian Penerapan Konsep Smart Environment sebagai Bagian dari Smart City (Studi Kasus: Kota Semarang),” 2024

[2] H. Sharfina, P. Y. Utami, and I. Fakhruzi, “Prediksi Bencana Banjir Menggunakan Algoritma Deep Learning H2O Berdasarkan Data Curah Hujan,” 2023. doi: https://doi.org/10.35957/jatisi.v10i4.5981.

[3] N. M. . Anggraeni, S. ., and Y. ., “Analisis Dampak Perubahan Iklim dan Pola Angin Pada Lingkungan Global”, *jpst*, vol. 2, no. 3, pp. 1041–1047, Dec. 2023. https://doi.org/10.47233/jpst.v2i4.1366.

[4] R. Afrian, “Kajian Mitigasi Terhadap Penyebab Bencana Banjir di Desa Sidodadi Kota Langsa,” 2021, doi: https://doi.org/10.32663/georaf.v5i2.1660.

[5] Wulandari, E. S. P., and Aziz, R. A., “Model Prediksi Dengan Artificial Neural Network Untuk Kejadian Banjir Rob di Wilayah Pesisir Kota Bandar Lampung,” 2022.

[6] I. Fitriyaningsih, Y. Basani, and L. M. Ginting, “MACHINE LEARNING: PROSPERITY OF RAINFALL, WATER DISCHARGE, AND FLOOD WITH WEB APPLICATION IN DELI SERDANG,” *JURNAL PENELITIAN KOMUNIKASI DAN OPINI PUBLIK*, vol. 22, no. 2, Dec. 2022, doi: 10.33299/jpkop.22.2.1752.

[7] S. Rizal, “Development of Big Data Analytics Model,” *ITEJ Juli-2019*, vol. 4, no. 1, 2019, doi: https://doi.org/10.24235/itej.v4i1.47.

[8] N. Yudistira, “Peran Big Data dan Deep Learning untuk Menyelesaikan Permasalahan Secara Komprehensif,” *EXPERT: Jurnal Manajemen Sistem Informasi dan Teknologi*, vol. 11, no. 2, p. 78, Dec. 2021, doi: 10.36448/expert.v11i2.2063.

[9] 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,” *TELKA*, vol. 9, no. 1, pp. 52–61, 2023.

[10] A. M. Siregar, “Klasifikasi Untuk Prediksi Cuaca Menggunakan Esemble Learning,” *PETIR*, vol. 13, no. 2, pp. 138–147, Sep. 2020, doi: 10.33322/petir.v13i2.998.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024 Page: 1222−1230
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5593

[11] E. Tangkelobo, W. Mayaut, H. Listanto, I. Binanto, and N. F. Sianipar, "Perbandingan Algoritma Klasifikasi Random Forest, Gaussian Naive Bayes, dan K-Nearest untuk Data Tidak Seimbang dan Data yang diseimbangkan dengan metode Random Undersampling pada dataset LCMS Tanaman Keladi Tikus," 2023. doi: https://doi.org/10.35842/sintaks.v2i1.28.

[12] M. Azhari, Z. Situmorang, and R. Rosnelly, "Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 2, p. 640, Apr. 2021, doi: 10.30865/mib.v5i2.2937.

[13] Hammam Riza, E. W. S. Santoso, Iwan Gunawan Tejakusuma, Firman Prawiradisastra, and Prihartanto, "Pemanfaatan Kecerdasan Artifisial untuk Meningkatkan Mitigasi Bencana Banjir," in *Prosiding Use Cases Artificial Intelligence Indonesia: Embracing Collaboration for Research and Industrial Innovation in Artificial Intelligence*, Penerbit BRIN, 2023. doi: 10.55981/brin.668.c545.

[14] I. Daniel, Z. Situmorang, J. Setia Budi, K. Tengah, and K. Medan Tuntungan, "Analysis of Machine Learning Algorithms in Predicting the Flood Status of Jakarta City," 2023. doi: https://doi.org/10.35842/icostec.v2i1.38.

[15] M. Putra, M. S. Rosid, and D. Handoko, "Rainfall Estimation Model in Seasonal Zone and Non-Seasonal Zone Regions Using Weather Radar Imagery Based on a Gradient Boosting Algorithm," *Atmosphere (Basel)*, vol. 15, no. 6, Jun. 2024, doi: 10.3390/atmos15060726.

[16] N. Hidayat, "Flood Disaster Detection Based on Rainfall Using Random Forest Algorithm," 2023. [Online]. Available: https://asasijournal.com/index.php/fcsj/article/view/15

[17] S. Dwiasnati and Yudo Devianto, "Optimization of Flood Prediction using SVM Algorithm to determine Flood Prone Areas," *Journal of Systems Engineering and Information Technology (JOSEIT)*, vol. 1, no. 2, pp. 40–46, Sep. 2022, doi: 10.29207/joseit.v1i2.1995.

[18] N. Fadhlina Mohd Anafi, N. Mohd Noor, H. Widyasamratri, and N. Mohn Noor, "A SYSTEMATIC REVIEW OF REAL-TIME URBAN FLOOD FORECASTING MODEL IN MALAYSIA AND INDONESIA-CURRENT MODELLING AND CHALLENGE," 2023.

[19] D. Feblian and D. U. Daihani, "IMPLEMENTASI MODEL CRISP-DM UNTUK MENENTUKAN SALES PIPELINE PADA PT X".2016.

[20] Y. A. Singgalen, "Analisis Performa Algoritma NBC, DT, SVM dalam Klasifikasi Data Ulasan Pengunjung Candi Borobudur Berbasis CRISP-DM," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 3, Dec. 2022, doi: 10.47065/bits.v4i3.2766.

[21] E. W. T. Ngai, L. Xiu, and D. C. K. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification," 2009, *Elsevier Ltd*. doi: 10.1016/j.eswa.2008.02.021.

[22] S. Siddique, M. A. Haque, R. George, K. D. Gupta, D. Gupta, and M. J. H. Faruk, "Survey on Machine Learning Biases and Mitigation Techniques," *Digital*, vol. 4, no. 1, pp. 1–68, Mar. 2024, doi: 10.3390/digital4010001.

[23] G. Mariscal, Ó. Marbán, and C. Fernández, "A survey of data mining and knowledge discovery process models and methodologies," Jun. 2010. doi: 10.1017/S0269888910000032.

[24] S. Singh and J. Prasad, "Estimation of Missing Values in the Data Mining and Comparison of Imputation Methods," *Mathematical Journal of Interdisciplinary Sciences*, vol. 1, no. 2, pp. 75–90, Mar. 2013, doi: 10.15415/mjis.2013.12015.

[25] A. Triayudi, Sumiati, T. Nurhadiyan H, and V. Rosalina, "Data mining implementation to predict sales using time series method," in *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, Institute of Advanced Engineering and Science, 2020, pp. 1–6. doi: 10.11591/eecsi.v7.2028.

[26] R. Alkentar and T. Mankovits, "Optimization of Additively Manufactured and Lattice-Structured Hip Implants Using the Linear Regression Algorithm from the Scikit-Learn Library," *Crystals (Basel)*, vol. 13, no. 10, Oct. 2023, doi: 10.3390/cryst13101513.

[27] A. M. Abdulazeez, M. A. Sulaiman, and D. Q. Zeebaree, "Evaluating Data Mining Classification Methods Performance in Internet of Things Applications," *Journal of Soft Computing and Data Mining*, vol. 1, no. 2, pp. 11–25, 2020, doi: 10.30880/jscdm.2020.01.02.002.