Seminar Nasional Sains Data 2024 (SENADA 2024)
UPN "Veteran" Jawa Timur
E-ISSN 2808-5841
P-ISSN 2808-7283

Penerapan *Machine Learning* dalam Klasifikasi Kejadian Hujan di Kabupaten Tuban Tahun 2019-2024

Adilla Khoirunnisa[1], Natasya Yunita Putri[2], Kenny Marsell Venezia Raiqhan[3], Robert Kurniawan[4]
[1, 2, 3]*Program Studi Statistika, Politeknik Statistika STIS, Jakarta, Indonesia*
[4]*Program Studi Komputasi Statistik, Politeknik Statistika STIS, Jakarta, Indonesia*
[2]212112244@stis.ac.id
[3]212112132@stis.ac.id
[4]robertk@stis.ac.id
Corresponding author email: 212111842@stis.ac.id

Abstract: *One of the impacts of climate change and the El Niño phenomenon is extreme drought. Tuban Regency, an important part of Indonesia's agricultural sector, is an area prone to drought disasters. The agricultural system in this region mostly uses a rain-fed rice field system with rainwater as the main water source. Therefore, the aim of this research is to find out the best model for classifying the occurrence of rain which determines the availability of water sources in rainfed rice fields in Tuban Regency. The data used is daily rainfall, average wind speed, highest wind direction, average humidity, duration of sunlight and average temperature taken from the official BMKG website. The methods used in this research include Decision Tree, Naïve Bayes, and SVM. Based on the results of the comparison of classification methods carried out, the best model is Naïve Bayes with an accuracy of 78.06 percent.*
Keywords: *Rainfall, Decision Tree, Naïve Bayes, SVM*

Abstrak: Salah satu dampak dari perubahan iklim dan fenomena *El Niño* adalah kekeringan ekstrem. Kabupaten Tuban, salah satu bagian penting dari sektor pertanian Indonesia, merupakan daerah yang rawan mengalami bencana kekeringan. Sistem pertanian pada wilayah ini sebagian besar menggunakan sistem sawah tadah hujan dengan air hujan sebagai sumber air utama. Karenanya, tujuan dari penelitian ini untuk mengetahui model terbaik dalam mengklasifikasikan terjadinya hujan yang menentukan ketersediaan sumber air pada pertanian sawah tadah hujan di Kabupaten Tuban. Data yang digunakan adalah curah hujan harian, kecepatan angin rata-rata, arah angin terbanyak, kelembapan rata-rata, lama penyinaran matahari, dan temperatur rata-rata yang diambil dari situs resmi BMKG. Metode yang digunakan dalam penelitian ini meliputi *Decision Tree, Naïve Bayes*, dan SVM. Berdasarkan hasil perbandingan metode klasifikasi yang dilakukan, model terbaik adalah *Naïve Bayes* dengan akurasi sebesar 78,06 persen.
Kata kunci: *Curah Hujan, Decision Tree, Naïve Bayes, SVM*

I. PENDAHULUAN

Indonesia berada di wilayah beriklim tropis yang menyebabkan wilayahnya hanya memiliki dua musim, yaitu musim hujan dan musim kemarau. Umumnya, musim kemarau di Indonesia berlangsung dari bulan April hingga bulan September, sementara musim penghujan berlangsung dari bulan Oktober hingga bulan Maret. Namun, adanya dampak perubahan iklim dan fenomena *El Niño* menyebabkan musim kemarau di Indonesia menjadi lebih panjang daripada biasanya. Tak hanya itu saja, 116 stasiun pengamatan BMKG juga mencatat rata-rata suhu udara pada tahun 2023 sebesar 27,2 °C, yang menyebabkan anomali rata-rata suhu udara pada tahun 2023 sebesar 0,5 °C dibandingkan rata-rata suhu udara pada periode tahun 1991-2020. Hal ini menjadikan tahun 2023 sebagai tahun terpanas ke-2 sepanjang periode pengamatan sejak tahun 1981 hingga 2023 di Indonesia. Bahkan, suhu udara yang telah tinggi ini terus mengalami peningkatan dan mencapai angka sebesar 27,74 °C pada bulan April 2024 lalu, menunjukan anomali positif dengan nilai sebesar 0,89 °C dibandingkan normal suhu udara klimatologis untuk bulan April 2024 periode 1991-2020 di Indonesia yang sebesar 26,85 °C. Anomali ini menjadi nilai tertinggi sepanjang periode pengamatan sejak tahun 1981.

Sebagian besar wilayah di Indonesia rentan terhadap dampak perubahan iklim seperti kenaikan suhu, pola hujan tak menentu, dan bencana alam seperti banjir dan kekeringan. Kondisi ini semakin

diperparah dengan adanya dampak dari fenomena *El Niño*. *El Niño* sendiri merupakan fenomena peningkatan suhu permukaan laut di Samudra Pasifik Tengah dan Timur yang dilepaskan ke atmosfer sehingga dapat mempengaruhi pola cuaca dan berpengaruh signifikan terhadap iklim di berbagai wilayah. *World Meteorological Organization* (WMO) menyatakan bahwa adanya suhu yang tinggi dapat berdampak pada kesehatan manusia dan kebakaran hutan di berbagai lokasi. Di Indonesia sendiri, fenomena *El Niño* ini telah memberikan dampak yang signifikan terhadap kekeringan, kekurangan air bersih, gagal panen, dan karhutla.

Jika melihat pada masa lalu, Indonesia pernah beberapa kali mengalami kekeringan terparah atau ekstrem, yaitu pada tahun 1997, 2015, dan 2019, yang mana penyebab utamanya ini dikarenakan fenomena *El Niño*. Kekeringan parah yang terjadi pada tahun 2019 berdampak pada sektor pertanian, sumber daya air, kehutanan, dan lingkungan, yang mana hal ini dipicu oleh adanya fenomena *El Niño* aktif sejak bulan September 2018 hingga bulan Juli 2019 di Samudera Pasifik ekuator bagian tengah. Kekeringan meteorologi ekstrem ini tetap terjadi, meskipun *El Niño* sudah kembali netral dan sebelumnya lemah [1]. Hal ini diikuti pula oleh fenomena *Dipole Mode* fase positif Samudera Hindia (IOD+) yang menguat sejak bulan April 2019 hingga bulan Desember 2019. Kekeringan parah tersebut juga disebabkan oleh suhu permukaan laut di wilayah Indonesia terutama pada bagian selatan yang lebih dingin (kurang dari 5 °C) daripada kondisi normalnya pada periode bulan Juni-November 2019 sehingga menyebabkan sukarnya pertumbuhan awan berpotensi hujan sebagai akibat dari kurangnya kadar uap air pada atmosfer yang dikarenakan rendahnya penguapan dari lautan. Musim kemarau di tahun 2019 pun cenderung lebih panjang daripada keadaan normalnya, dimana kurang lebih 46% dari 342 zona musim di Indonesia mengalami musim kemarau dengan durasi yang sama hingga 6 dasarian (2 bulan) lebih panjang.

Kekeringan, terlebih kekeringan parah, dengan pola curah hujan tak pasti dan kenaikan suhu yang ekstrem tentu akan dapat memberikan kerugian besar di berbagai sektor kehidupan manusia, terutama pada sektor pertanian yang dapat berdampak pada ketersediaan pangan masyarakat. Dalam kondisi kemarau berkepanjangan dengan intensitas curah hujan yang rendah, tanaman rawan mengalami stres panas dan lahan pertanian pun dapat mengalami kekeringan yang dapat mengurangi produktivitas pertanian dan menekan hasil panen. Bahkan, dikhawatirkan hal ini dapat berdampak lebih parah dengan berujung pada kegagalan panen, khususnya pada tanaman pangan semusim yang sangat mengandalkan air. Curah hujan tak pasti dan kekeringan berkepanjangan akan menyebabkan sulitnya pengelolaan tanaman dan ternak, juga meningkatkan risiko terjadinya kebakaran hutan yang dapat merusak lahan pertanian. Selain itu, ketersediaan air untuk irigasi yang sangat penting dalam pertanian juga dapat dipengaruhi oleh kondisi kekeringan ini. Tentu saja berbagai situasi tersebut dapat mengarah dan menjadi ancaman serius terhadap kondisi ketahanan pangan lokal.

Jawa Timur merupakan wilayah dengan produksi panen padi terbesar se-Indonesia. Wilayah ini memiliki karakteristik curah hujan yang lebih rendah secara alami dan musim kemarau yang panjang [2], yang menyebabkan wilayah ini juga turut masuk ke dalam salah satu daerah terdampak kekeringan parah pada tahun 2019 silam. Menurut Dinas Pertanian Provinsi Jawa Timur, luas lahan pertanian yang terkena dampak kekeringan di Jawa Timur pada tahun 2019 seluas 24.633 hektar dan luas puso sekitar 983 hektar. Salah satu kabupatennya, yaitu kabupaten Tuban, merupakan bagian penting dari sektor pertanian di Indonesia karena wilayah ini termasuk ke dalam lima besar daerah dengan produksi panen padi terbanyak se-Jawa Timur. Namun, Tuban sendiri sebenarnya merupakan salah satu daerah yang rawan mengalami bencana kekeringan. Pada tahun 2023, BNPB (Badan Nasional Penanggulangan Bencana) merilis skor IRBI Kabupaten Tuban sebesar 15,53 untuk kategori resiko bencana kekeringan,

yang mana nilai masuk dalam kategori tingkat resiko tinggi untuk bencana kekeringan [3]. Daerah Tuban pun juga turut merasakan dampak dari kekeringan parah pada tahun 2019 lalu, dimana luas lahan pertanian yang terkena dampak kekeringan seluas 2.768 hektar. Wilayah kekeringan ini terbagi dalam beberapa kategori, yaitu kategori ringan seluas 1.453,1 hektar, kategori sedang seluas 783,2 hektar, kategori berat 512 hektar, dan puso atau gagal panen seluas 65 hektar [4].

Kabupaten Tuban merupakan salah satu daerah kering dalam pertanian lahan kering, tidak banyak jenis tanaman yang tumbuh di lingkungan tersebut karena kurangnya air dan unsur hara yang dimiliki oleh tanah. Karenanya, menurut [5], pertanian yang ada di wilayah Tuban sangat bergantung pada curah hujan, hal ini disebabkan oleh sebagian besar pertaniannya yang menggunakan sistem pertanian sawah tadah hujan, yang mana sumber air utamanya berasal dari air hujan. Jika tidak ada curah hujan, kelembaban tanah menjadi rendah sehingga tidak menguntungkan untuk budidaya padi [2].

Beberapa penelitian telah mengkaji faktor-faktor yang dapat mempengaruhi pertumbuhan dan produktivitas padi, terutama pada curah hujan. Penelitian oleh [6] menyatakan bahwa suhu maksimum berdampak negatif terhadap tanaman padi sehingga mengakibatkan penurunan jumlah tanaman pada tahap penanaman kembali. Dampak positif dari suhu minimum terhadap produksi padi juga diamati yang dapat menyebabkan pertumbuhan tanaman, yang mempengaruhi tanaman padi pada tahap penanaman kembali selama fase vegetatif. Hasil penelitian menunjukkan bahwa jumlah anakan dan pola makan tanaman padi meningkat seiring dengan dampak positif curah hujan pada tahap anakan. Suhu maksimum mempunyai dampak negatif terhadap tanaman padi pada tahap anakan dan pemanjangan batang. Telah diamati bahwa curah hujan mempunyai dampak negatif pada tanaman padi pada tahap awal dan masa pembungaan. Penurunan produksi padi secara signifikan terjadi karena rusaknya sel-sel reproduksi pada tahap menuju dan berhenti selama fase reproduksi. Dampak negatif curah hujan terhadap produksi padi juga terlihat pada tahap pemerahan selama fase pemasakan. Dampak negatif yang signifikan dari curah hujan terlihat pada produk domestik bruto per kapita selama fase reproduksi.

Selain itu, penelitian oleh [7] juga menyatakan pengaruh dari angin dan curah hujan yang dapat mengurangi produksi padi, yang mana pengaruhnya tergantung pada genotipe dan tahap pembungaan. Penelitian [8] menemukan bahwa prediksi hasil panen padi dan hasil nyata mempunyai hubungan nyata dengan curah hujan, suhu maksimum dan suhu minimum. Penelitian oleh [9] mengemukakan bahwa seiring dengan kurangnya curah hujan, kelembaban tanah dan suhu di atas normal (terutama suhu maksimum) yang terjadi secara terus-menerus ditemukan menjadi faktor penting dalam kasus hilangnya produktivitas padi secara parah.Hujan merupakan sebuah siklus hidrologi yang dimulai dari proses penguapan air yang terjadi di permukaan bumi, baik dari wilayah perairan, tumbuhan, maupun daratan, kemudian mengalami proses kondensasi uap air di atmosfer sehingga menjadi butiran air dan akhirnya terjadilah proses presipitasi uap air dari awan menuju ke permukaan bumi kembali. Berdasar hal tersebut, dapat dikatakan bahwa peristiwa terjadinya hujan ini melalui beberapa proses siklus air. Siklus hidrologi ini menggambarkan penyimpanan dan pergerakan air secara terus menerus antara biosfer, atmosfer, kriosfer, litosfer, antroposfer, dan hidrosfer [10].

Secara umum, tahapan terjadinya hujan terbagi menjadi tiga, yaitu evaporasi, kondensasi, dan presipitasi. Tahapan pertama, yaitu tahap evaporasi, merupakan proses perubahan wujud dari air menjadi uap atau biasa disebut proses penguapan air. Panasnya suhu bumi karena penyinaran matahari akan mengakibatkan air-air di permukaan bumi mengalami penguapan. Semakin panas suhu udara dan semakin lama penyinaran matahari di suatu wilayah, maka akan semakin banyak pula air yang menguap ke udara dan naik ke atmosfer. Tahapan evaporasi ini mengalirkan air ke atmosfer dengan kecepatan

yang bervariasi sesuai dengan kondisi iklim. Selanjutnya, uap air yang telah naik cukup tinggi ke atmosfer akan mengalami proses kondensasi atau pengembunan. Pada tahapan ini, uap air akan memadat menjadi partikel-partikel es yang sangat kecil. Hal ini bergantung pada suhu udara dan suhu titik embun. Suhu titik embun sendiri adalah suhu dimana udara, ketika didinginkan, menjadi jenuh dan embun dapat terbentuk. Ketika suhu udara dan suhu titik embun sama, maka terjadilah kabut. Karena uap air memiliki tingkat energi yang lebih tinggi daripada air, ketika terjadi kondensasi, kelebihan energi tersebut dilepaskan dalam bentuk panas. Ketika partikel-partikel es yang memiliki jari-jari sekitar 5-20 mm saling berdekatan atau bertumbuk dan bergabung satu sama lain, maka terbentuklah gumpalan awan, yang mana proses pembentukan awan ini dikenal dengan istilah koalesensi. Dengan ukuran yang sangat kecil tersebut, partikel akan jatuh dalam kecepatan 0,01–5 cm/s. Namun, bila kecepatan aliran udara lebih tinggi, maka partikel tersebut tidak akan jatuh ke permukaan bumi.

Adanya perbedaan suhu dan ketinggian awan di atmosfer akan mempengaruhi proses perubahan uap air menjadi partikel es tersebut. Jika keberadaan awan di atmosfer semakin tinggi, maka akan menyebabkan suhu udara semakin dingin. Namun, perlu diingat bahwa tidak semua air yang melalui proses kondensasi akan membentuk awan. Bisa saja sebagian mengembun di dekat tanah, sebagian naik menjadi kabut, sementara sebagian sisanya baru menguap ke atmosfer membentuk awan. Lalu, tahapan yang terakhir adalah proses presipitasi. Presipitasi merupakan peristiwa jatuhnya partikel es yang telah mencair dari atmosfer dan turun menjadi tetesan-tetesan hujan ke permukaan bumi. Ketika awan telah mencapai tingkat kepadatan uap air yang maksimal dan tidak mampu menahan beban airnya lagi, maka awan tersebut akan membentuk titik-titik hujan dan jatuh ke permukaan bumi. Akibat terbawa angin yang bergerak, awan yang telah terbentuk sebelumnya dapat berpindah dan menyebabkan terjadinya hujan di tempat yang berbeda dari proses siklus air sebelumnya. Berdasarkan proses pembentukan hujan yang telah diuraikan di atas, curah hujan dapat terbentuk dengan saling berkaitan dengan variabel-variabel lain, seperti suhu atau temperatur udara, penyinaran matahari, kecepatan dan arah angin, serta kelembaban udara. Beberapa penelitian terdahulu telah melakukan pengkajian dan pengklasifikasian hubungan antara variabel-variabel tersebut dengan berbagai metode klasifikasi.

Penelitian yang dilakukan oleh [11] memperkirakan lokasi yang berpotensi hujan dengan mempertimbangkan berbagai faktor, seperti kelembapan, kecepatan angin, ketinggian air, dan suhu menggunakan berbagai metode mencakup *Naïve Bayes*, *Decision Tree*, *Support Vector Machine (SVM)*, *Random Forest*, dan *Neural Network*. Sementara, [12] menganalisis berbagai algoritma *machine learning* untuk memprediksi curah hujan dengan fitur yang dipilih sebagai variabel masukan, diantaranya suhu maksimum, suhu minimum, sinar matahari, kecepatan angin, kelembapan, dan penguapan. Penelitian lain oleh [13], menghasilkan penentuan keakuratan curah hujan menggunakan metode *Naïve Bayes Classifier* (NBC) dengan menggunakan beberapa parameter, seperti rata-rata kelembapan udara dan rata-rata kecepatan angin. Selanjutnya, juga ada penelitian oleh [14] yang melakukan perbandingan kinerja *Random Forest*, *Decision Tree*, dan *Support Vector Machine* untuk prediksi curah hujan dengan suhu minimum dan suhu maksimum mengukur data atribut terkait curah hujan harian. Penelitian [15] melakukan prediksi curah hujan dengan didasarkan pada mesin vektor pendukung (SVM). Hasil penelitian yang dilakukan [16] mengulas algoritma pembelajaran ansambel, antara lain *bagging*, *boosting*, dan *stacking* untuk memprediksi curah hujan.

Pentingnya curah hujan dalam sistem pertanian di wilayah Tuban menarik minat peneliti untuk melakukan pengklasifikasian terhadap kejadian hujan harian di wilayah Tuban. Pengklasifikasian dilakukan dengan membandingkan berbagai metode klasifikasi untuk mendapatkan model dengan akurasi terbaik sehingga prediksi yang akan dilakukan dapat memberikan hasil yang lebih tepat.

Penelitian lain mengklasifikasikan curah hujan di kabupaten Banyuwangi menjadi hujan ringan, normal, dan lebat dengan menggunakan metode *Naïve Bayes*, yang bertujuan untuk melakukan prediksi karena pertanian dan perkebunannya cenderung terletak di daerah terpencil yang cenderung kekurangan informasi terkait cuaca dan iklim. Namun, penelitian ini tidak ikut memperhitungkan kategori tidak hujan, padahal kategori ini penting untuk mengetahui kemungkinan hari tanpa hujan yang dapat berdampak pada keadaan air tanah di pertanian. Karena stres air tanah pada tahap vegetatif dan generatif tanaman padi mengurangi pertumbuhan tanaman dan hasil panen, serta mengurangi efisiensi penggunaan air tanaman [17]. Selain itu, penelitian tersebut juga hanya menggunakan 2 variabel masukan, kurang dari jumlah kategori yang digunakan, yaitu sebanyak 3. Oleh karenanya, penelitian ini bertujuan untuk melakukan pengklasifikasian hujan dan tidak hujan sebagai penentu ketersediaan air pada pertanian sawah tadah hujan di Kabupaten Tuban menggunakan lima variabel masukan, dengan harapannya penelitian ini dapat berkontribusi pada petani dalam menyiapkan pengelolaan sumber air pertanian.

II. METODE PENELITIAN

II.1. Sumber Data

Pada penelitian ini, data yang digunakan adalah data sekunder runtun waktu harian dengan rentang bulan Januari 2019 hingga April 2024 yang bersumber dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG). Lokus yang diteliti dalam penelitian ini adalah Kabupaten Tuban, Jawa Timur dengan total 1947 amatan. Adapun variabel yang digunakan pada penelitian ini adalah sebagai berikut.

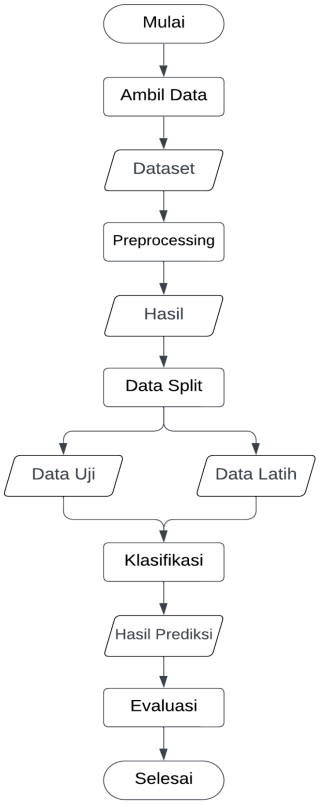
Tabel 1. Atribut Data

Variabel	Definisi	Satuan	Sumber
Curah Hujan	Ketinggian air hujan yang ditampung oleh alat pengukur hujan yang ditempatkan pada tempat yang datar dan kedap air	mm/hari	BMKG
Kecepatan Angin Rata-rata	Nilai rata-rata kecepatan aliran udara dari tekanan tinggi ke tekanan rendah dalam periode waktu tertentu	m/s	BMKG
Arah Angin Terbanyak	Arah terbanyak dari mana angin berembus dalam periode waktu tertentu yang ditentukan searah jarum jam dan dimulai dari titik utara bumi	^{O}	BMKG
Kelembapan Rata-rata	Nilai rata-rata besarnya kandungan uap air yang dikandung oleh udara dalam periode waktu tertentu	persen	BMKG
Lama Penyinaran Matahari	Lamanya waktu sinar matahari menyinari permukaan bumi dalam periode satu hari	jam	BMKG
Temperatur Rata-rata	Nilai rata-rata suhu di suatu lokasi atau wilayah dalam periode waktu tertentu	^{O}C	BMKG

II.2. Teknik Analisis Data

Analisis data dilakukan dengan menggunakan beberapa metode klasifikasi *data mining*, yaitu *Decision Tree*, *Naïve Bayes*, dan SVM, yang kemudian akan diputuskan metode terbaik dalam mengklasifikasikan data. Dalam melakukan analisis, pengolahan data ini dibantu *software* berbasis *data mining*, yaitu Rstudio. Adapun alur penelitian dapat dilihat pada diagram alir berikut ini.

E-ISSN 2808-5841
P-ISSN 2808-7283



Gambar 1. Alur Penelitian

Tahapan penerapan metode klasifikasi ini secara berturut-turut adalah sebagai berikut.

- Melakukan Pengambilan Data
  Data yang digunakan pada penelitian ini bersumber dari laman peramban resmi BMKG dengan total lima variabel independen dan satu variabel dependen. Data yang telah diunduh dari laman kemudian dijadikan satu menjadi sebuah *dataset*.
- Melakukan *Data Preprocessing*
  Preprocessing atau pemrosesan awal data adalah serangkaian teknik yang digunakan sebelum menerapkan metode *data mining*. Secara umum, *data preprocessing* meliputi tahapan pembersihan, penggabungan, dan transformasi data agar data dapat dilakukan pengolahan lebih lanjut [18].
- Melakukan *Data Split*
  Data split atau pemisahan data adalah teknik membagi data menjadi dua, yaitu data uji dan data latih, yang bertujuan untuk mengurangi atau menghilangkan bias terhadap data latih dan mencegah algoritma menghasilkan *overfitting* yang berdampak negatif pada data uji yang sebenarnya [19].
- Melakukan Klasifikasi Data
  Metode klasifikasi *data mining* yang digunakan pada penelitian ini adalah *Decision Tree, Naïve Bayes,* dan SVM.

1. *Decision Tree*
   Metode *decision tree* adalah algoritma berbasis pohon dengan teknik pengelompokan data yang tersusun dengan menggunakan aturan untuk memisahkan data hingga mencapai hasil akhir yang diinginkan [20].

2. *Naïve Bayes*
   Metode *Naïve Bayes* adalah algoritma yang dapat digunakan untuk memprediksi probabilitas keanggotaan kelas, dapat diandalkan untuk menangani kumpulan data yang besar, dan mampu mengatasi data yang tidak relevan [13].
3. SVM
   Metode SVM adalah algoritma dengan dasar teori yang solid dan optimalisasi global yang dirancang untuk klasifikasi biner serta membutuhkan pendekatan berbeda untuk masalah *multilevel* yang dipecahkan menjadi serangkaian masalah klasifikasi biner ganda [21, 22].

- Melakukan Evaluasi Data
  Evaluasi data dilakukan untuk memilih model terbaik dari metode yang sebelumnya digunakan dalam mengklasifikasikan data dengan menggunakan *confusion matrix*. *Confusion matrix* adalah suatu tabel persegi dengan baris mewakili kelas aktual dan kolom mewakili prediksi model yang digunakan untuk mengevaluasi kinerja model dalam klasifikasi [23]. Dalam klasifikasi biner, *confusion matrix* memiliki ukuran 2x2 yang memberi informasi mengenai jumlah prediksi benar dan salah. Berikut adalah gambaran mengenai *confusion matrix*.

Tabel 2. *Confusion Matrix*

	Predicted	
Actual Class	*C*	*-C*
C	*TP*	*FN*
-C	*FP*	*TN*

di mana
TP : prediksi benar (*event* terjadi)
FP : prediksi salah (*event* terjadi)
FN : prediksi salah (*event* tidak terjadi)
TN : prediksi benar (*event* tidak terjadi)

Dalam *confusion matrix*, terdapat nilai-nilai acuan yang digunakan untuk membandingkan metode, yaitu nilai akurasi, presisi, *recall*, dan *F-1 score* yang juga dihasilkan dari masing-masing metode. Pengertian dan perhitungan dari masing-masing acuan adalah sebagai berikut:

1. Akurasi adalah persentase kebenaran prediksi
   $\text{Akurasi} = \frac{(TP + TN)}{(TP + FP + FN + TN)}$;
2. Presisi adalah tingkat ketepatan hasil prediksi
   $\text{Presisi} = \frac{TP}{(TP + FP)}$;
3. *Recall* adalah tingkat sensitivitas terhadap bagian yang relevan
   $Recall = \frac{TP}{(TP + FN)}$;
4. *Sensitivity* adalah ukuran untuk mengukur seberapa baik suatu uji dalam menghindari hasil salah negatif
   $Sensitivity = \frac{TP}{P}$;
5. *Specificity* adalah ukuran untuk mengukur seberapa baik suatu uji dalam menghindari hasil salah positif
   $Specificity = \frac{TN}{N}$;
6. *F-1 Score* adalah rata-rata harmoni presisi dan *recall*

$$F\text{-}1\ Score = \frac{2\,(Recall * Presisi)}{(Recall + Presisi)}.$$

Akurasi direkomendasikan untuk digunakan sebagai acuan informasi jika *dataset* memiliki jumlah data FN dan FP yang simetris (jumlah hampir sama). Jika tidak simetris (FN dan FP berbeda jauh), maka menggunakan *F-1 Score* sebagai gantinya. *Recall* dapat digunakan sebagai acuan jika menginginkan FP untuk terjadi daripada FN. Jika lebih menginginkan untuk terjadi TP namun tidak menginginkan terjadi FP, maka acuan presisi dapat digunakan. Setelah meninjau dan memilih acuan, maka model yang memiliki nilai tertinggi pada acuan-acuan tadi akan ditetapkan sebagai metode terbaik.

III. HASIL DAN PEMBAHASAN

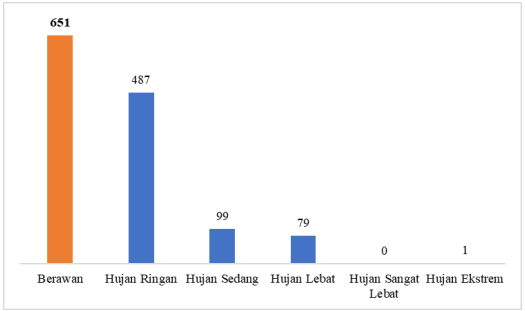
III.1. Data Preprocessing

Klasifikasi curah hujan didasarkan dari nilai ambang BMKG yang disajikan pada tabel 3. Untuk memenuhi tujuan penelitian, dilakukan pengkategorian ulang curah hujan dengan dua kategori, yaitu hujan dan tidak hujan. Intensitas curah hujan yang dikategorikan hujan meliputi hujan ringan, hujan sedang, hujan lebat, dan hujan sangat lebat, sedangkan berawan dikategorikan tidak hujan dengan *cut off* sebesar 0,5 mm/hari.

Tabel 3. Kategori Hujan BMKG

No.	Kategori	Intensitas Curah Hujan (mm/hari)
1.	Berawan	0 mm/hari
2.	Hujan Ringan	0,5 - 20 mm/hari
3.	Hujan Sedang	20-50 mm/hari
4.	Hujan Lebat	50-100 mm/hari
5.	Hujan Sangat Lebat	100-150 mm/hari
6.	Hujan Ekstrem	>150 mm/hari

Pada variabel curah hujan, terdapat 158 dari 1947 data berkode 8888 yang merupakan data tidak terukur. Selain itu, terdapat 472 missing value. Untuk itu, pada tahap Data Cleaning, dilakukan pembersihan dengan menghilangkan data berkode 8888 dan missing value, sehingga hanya tersisa 1317 data.



Gambar 2. Kejadian Hujan Berdasarkan Kategori BMKG di Kabupaten Tuban

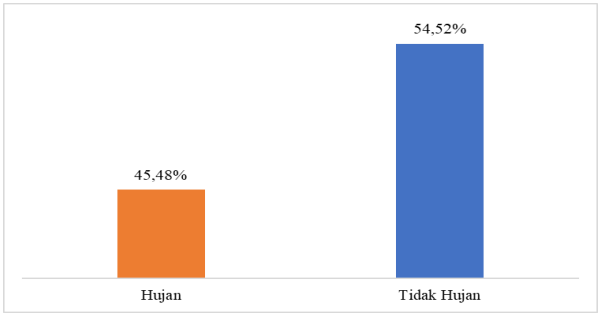
Gambar di atas menunjukkan banyaknya kejadian hujan berdasarkan kategori hujan BMKG dari hasil *Data Cleaning*. Kategori berawan mendominasi kejadian hujan di Kabupaten Tuban, dilanjutkan

dengan hujan ringan, hujan sedang, dan hujan lebat. Di sisi lain, kejadian hujan ekstrem sangat jarang terjadi, sedangkan hujan sangat lebat tidak pernah terjadi dalam rentang waktu yang diamati. Selanjutnya, dilakukan diskretisasi untuk mengubah data kontinyu menjadi data numerik. Hal ini perlu dilakukan agar syarat penggunaan metode klasifikasi dapat terpenuhi. Diskretisasi dilakukan pada variabel temperatur rata-rata dan lama penyinaran matahari. Pada tahap selanjutnya, kategori hujan diberi label angka 1 (satu), sedangkan tidak hujan diberi label angka 0 (nol). Tabel 4 menunjukkan hasil *data preprocessing* yang akan digunakan untuk tahapan selanjutnya.

Tabel 4. Data

Tanggal	Temperatur Rata-Rata (^{O}C)	Kelembapan Rata-Rata (%)	Lama Penyinaran Matahari (Jam)	Kecepatan Angin Rata-Rata (m/s)	Arah Angin Terbanyak	Kejadian Hujan
1/1/2019	27	85	3	4	West	1
1/4/2019	27	87	0	1	Calm	1
1/5/2019	28	80	5	2	Calm	0
1/7/2019	27	85	10	2	Calm	1
1/9/2019	28	84	10	2	South West	1
				…..		
4/29/2024	29	81	6	1	Calm	1
4/30/2024	29	81	9	2	East	1

Gambaran secara umum proporsi hujan dan tidak hujan hasil *preprocessing* di Kabupaten Tuban dapat dilihat pada Gambar 3.



Gambar 3. Proporsi Kejadian Hujan dan Tidak Hujan di Kabupaten Tuban

Dari 1317 observasi, diperoleh bahwa proporsi kejadian tidak hujan sebanyak 718 atau 54,25 persen yang nilainya lebih tinggi dibandingkan proporsi hujan, yaitu sebanyak 599 atau 45,48 persen.

III.2. Splitting Data

Proses klasifikasi diawali dengan pembagian data menjadi data latih (*training*) dan data uji (*testing*). Tabel 5 menyajikan perbandingan rasio pembagian data berdasarkan nilai akurasi (%), presisi(%), *recall*(%), dan *F-1 score*(%).

Tabel 5. Perbandingan Rasio Pembagian Data Setiap Metode

Rasio	Parameter	Decision Tree	Naïve Bayes	SVM
70:30	Akurasi	74,11%	73,60%	74,11%
	Presisi	79,39%	71,19%	74,52%
	Recall	58,10%	70,39%	65,36%
	F1-Score	67,09%	70,79%	69,64%
75:25	Akurasi	73,17%	71,95%	74,09%
	Presisi	68,05%	67,92%	72,86%
	Recall	77,18%	72,48%	68,46%
	F1-Score	72,33%	70,13%	70,59%
80:20	Akurasi	74,43%	73,66%	75,19%
	Presisi	78,26%	70,49%	75,47%
	Recall	60,50%	72,27%	67,23%
	F1-Score	68,25%	71,37%	71,11%
85:15	Akurasi	75,51%	78,06%	77,55%
	Presisi	71,58%	77,38%	80,82%
	Recall	76,40%	73,03%	66,29%
	F1-Score	73,91%	75,14%	72,84%
90:10	Akurasi	70,00%	72,31%	73,08%
	Presisi	76,32%	69,49%	73,08%
	Recall	49,15%	69,49%	64,41%
	F1-Score	59,79%	69,49%	68,47%

Berdasarkan hasil perbandingan di atas, diperoleh nilai akurasi dan presisi tertinggi untuk semua model terdapat pada rasio 85:15, artinya data *training* sebesar 85 persen dan data *testing* sebesar 15 persen dari keseluruhan data. Oleh karena itu, rasio perbandingan data yang digunakan dalam penelitian ini adalah 85:15.

III.3. Perbandingan Performa Model Decision Tree, Naïve Bayes, dan SVM

Pemilihan model terbaik dilakukan dengan membandingkan nilai akurasi, presisi, *recall*, dan *F1-score*. Perbandingan ini tersaji dalam tabel 6.

Tabel 6. Perbandingan Performa Model *Decision Tree*, *Naïve Bayes*, dan SVM

Parameter	*Decision Tree*	*Naïve Bayes*	SVM
Akurasi	74,49%	78,06%	77,55%
Presisi	80,95%	77,38%	80,82%
Recall	57,30%	73,03%	66,29%
F1-Score	67,11%	75,14%	72,84%

Pada penelitian ini, akurasi berupa persentase observasi yang benar diprediksi sebagai hujan dan tidak hujan dari keseluruhan observasi. Algoritma dengan akurasi tertinggi adalah *Naïve Bayes* dengan nilai sebesar 78,06 persen. Ukuran selanjutnya merupakan presisi. Dalam penelitian ini, presisi berupa persentase banyaknya observasi yang benar merupakan hujan dari keseluruhan observasi yang diprediksi hujan. Algoritma dengan presisi tertinggi adalah *Decision Tree* sebesar 80,95 persen. Ukuran lain berupa *recall*, yaitu persentase observasi yang diprediksi hujan dari keseluruhan observasi yang sebenarnya hujan. *Naïve Bayes* adalah algoritma dengan *recall* tertinggi, yaitu sebesar 73,03 persen. Ukuran terakhir adalah *F1-Score* yang memuat perbandingan rata-rata presisi dan *recall* yang dibobotkan. Algoritma dengan *F1-Score* tertinggi adalah *Naïve Bayes* dengan nilai sebesar 75, 14 persen.

Naïve Bayes memiliki dua ukuran performa model yang lebih tinggi dibandingkan model lain, yaitu akurasi dan *F-1 Score*. Akurasi tertinggi cocok dipilih karena dapat menggambarkan keberhasilan prediksi secara umum, sedangkan *F-1 score* tertinggi dipilih karena telah mempertimbangkan keseimbangan antara presisi dan *recall*. Dengan demikian, model terbaik yang dipilih adalah *Naïve Bayes* dengan akurasi 78,06 persen. Setelah proses evaluasi model dan didapatkan model yang terbaik, dilakukan proses pengecekan *overfitting*. *Overfitting* merupakan kondisi model klasifikasi memiliki performa yang baik pada data *training*, tetapi tidak dengan data *testing. Overfitting* terjadi ketika akurasi dalam data *training* lebih tinggi dibandingkan akurasi data *testing*. Berdasarkan hasil pengecekan, diperoleh akurasi menggunakan data *training* sebesar 74,75 persen yang lebih kecil dibandingkan akurasi menggunakan data *testing*, yaitu sebesar 78,06 persen. Dengan demikian, pada model *Naïve Bayes* tidak terjadi *overfitting*, sehingga model ini dapat mengklasifikasikan data baru dengan baik. Hasil ini sesuai dengan penelitian yang dilakukan oleh [12] tentang klasifikasi curah hujan menjadi hujan ringan, normal, dan lebat dengan menggunakan metode *Naïve Bayes*. Penelitian tersebut menunjukkan bahwa model dapat mengklasifikasikan dengan baik, mencapai tingkat akurasi sebesar 95,92 persen. Penelitian lain [24] terkait prediksi potensi hujan di Kota Ternate menunjukkan hal yang sama, yang mana *Naïve Bayes* mampu mengklasifikasikan secara akurat dengan akurasi sebesar 76 persen.

III.4. Confusion Matrix Naïve Bayes

Tabel 7. *Confusion Matrix Naïve Bayes*

Observed		*Predicted*	
		Curah Hujan	
		Hujan	Tidak Hujan
Curah Hujan	Hujan	65	24
	Tidak Hujan	19	88
	Akurasi		78,06%

Berdasarkan tabel 7 tersebut, dapat dilihat bahwa 65 dari 89 data kategori "hujan" diprediksi benar, sedangkan 24 data diprediksi "tidak hujan". Selain itu, 88 dari 107 data kategori "tidak hujan" diprediksi benar, sedangkan 19 data diprediksi "hujan". Dengan demikian, nilai *sensitivity* sebesar 73,03 persen, artinya model memiliki kemampuan untuk mengidentifikasi sekitar 73,03 persen dari semua kejadian hujan yang sebenarnya, sedangkan nilai *specificity* sebesar 82,24 persen, artinya model memiliki kemampuan untuk mengidentifikasi sekitar 82,24 persen dari semua kejadian tidak hujan yang sebenarnya. Untuk itu, akurasi yang diperoleh sebesar 78,06 persen dengan *error rate* sebesar 21,94 persen, artinya dari seluruh data *testing* yang dievaluasi, 78,06 persen prediksi yang dibuat oleh model adalah benar, sedangkan 21,94 persen lainnya salah.

IV. KESIMPULAN

Hasil perbandingan metode klasifikasi menunjukkan bahwa *Naïve Bayes* menjadi model terbaik berdasarkan nilai akurasi dan *F-1 Score* teratas, sehingga mampu mengklasifikasikan hujan secara akurat di Kabupaten Tuban dengan nilai akurasi sebesar 78,06 persen. Untuk peneliti selanjutnya dapat melakukan pengujian ulang dengan metode lain atau menggabungkan beberapa metode dan menambah jumlah amatan dengan harapan dapat meningkatkan akurasi serta menutup kekurangan metode yang telah digunakan sebelumnya.

REFERENSI

1. S. Siswanto, K. K. Wardani, B. Purbantoro, A. Rustanto, F. Zulkarnain, E. Anggraheni, R. Dewanti, T. Nurlambang, M. Dimyati, "Satellite-Based Meteorological Drought Indicator to Support Food Security in Java Island," *PLoS One*, Vol. 17, No. 6 June, h. 1–20, 2022, doi: 10.1371/journal.pone.0260982.
2. H. Mulyanti, I. Istadi, and R. Gernowo, "Historical, Recent, and Future Threat of Drought on Agriculture in East Java, Indonesia: A Review," *E3S Web of Conferences*, Vol. 448, 2023, doi: 10.1051/e3sconf/202344803016.
3. BNPB, "Indeks Risiko Bencana Indonesia Tahun 2023", BNPB, Vol. 2, No. 2, 2023.
4. I. T. Amir, N. H. I. Fitriana, and E. Mulyana, "Risk Analysis of Dry Land Rice Production on the Impact of Climate Change and Weather," *Proceedings of the 3rd International Conference on Agriculture*, Vol. 1, 2023, doi: 10.2991/978-94-6463-168-5_15.
5. K. Putra, W. Baskoro, N. Ratini, N. Wendri, and A. Widagda, "Assessment of Groundwater Availability for Rice Farming in Tuban Regency, East Java in 2018-2022," *International Journal of Environtment and Climate Change*, Vol. 14, No. 2, h. 828–836, 2024, doi: 10.9734/ijecc/2024/v14i23995.
6. S. Abbas dan Z. A. Mayo, "Impact of Temperature and Rainfall on Rice Production in Punjab, Pakistan," *Environtment, Development and Sustainability*, Vol. 23, No. 2, h. 1706–1728, 2021, doi: 10.1007/s10668-020-00647-8.
7. H. Agusta, E. Santosa, Dulbari, D. Guntoro, and S. Zaman, "Continuous Heavy Rainfall and Wind Velocity During Flowering Affect Rice Production," *Agrivita Journal of Agricultural Science*, Vol. 44, No. 2, h. 290–302, 2022, doi: 10.17503/agrivita.v44i2.2539.
8. L. Wickramasinghe, J. Jayasinghe and U. Rathnayake, "Relationships between Climatic Factors to the Paddy Yield in the North-Western Province of Sri Lanka," *2020 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, pp. 223-227, 2020, doi: 10.1109/SCSE49731.2020.9313047.
9. N. Sahu, A. Saini, S. Behera, T. Sayama, S. Nayak, L. Sahu, W. Duan, R. Avtar, M. Yamada, R. B. Singh, K. Takara, "Impact of Indo-Pacific Climate Variability on Rice Productivity in Bihar, India," *Sustainability*, Vol. 12, No. 17, h. 1–21, 2020, doi: 10.3390/su12177023.
10. C. G. Collier, "Hydrometeorology (Advancing Weather and Climate Science)", 2016.
11. N. S. Sani, A. H. A. Rahman, A. Adam, I. Shlash, and M. Aliff, "Ensemble Learning for Rainfall Prediction," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 11, pp. 153–162, 2020, doi: 10.14569/IJACSA.2020.0111120.

12. C. M. Liyew, H. A. Melese, "Machine Learning Techniques to Predict Daily Rainfall Amount," *Journal of Big Data*, Vol. 8, No. 1, 2021, doi: 10.1186/s40537-021-00545-4.
13. A. U. Azmi, A. F. Hadi, D. Anggraeni, A. Riski, "Naïve Bayes Methods for Rainfall Prediction Classification in Banyuwangi," *Journal of Physics: Conference Series*, h. 1-8, 2020.
14. N. Paudel, T. N. Yogi, "Comparative Study Of Machine Learning Algorithms for Rainfall Prediction – A Case Study in Nepal," *International Journal of Advanced Research in Engineering and Technology*, Vol. 3, No. 3, h. 677–681, 2019, doi: 10.31142/ijtsrd22961.
15. E. Hussein, M. Ghaziasgar, and C. Thron, "Regional Rainfall Prediction Using Support Vector Machine Classification of Large-Scale Precipitation Maps," *2020 IEEE 23rd International Conference on Information Fusion*, h. 1-8, 2020, doi: 10.23919/FUSION45008.2020.9190285.
16. S. Kundu, S. K. Biswas, D. Tripathi, R. Karmakar, S. Majumdar, S. Mandal, "A Review on Rainfall Forecasting using Ensemble Learning Techniques," *e-Prime - Advances Electrical Engineering, Electroics and Energy*, Vol. 6, 2023, doi: 10.1016/j.prime.2023.100296.
17. Abdourasmane Kadougoudiou Konaté, Adama Zongo, Jean Rodrigue Sangaré, Audrey Dardou, and Alain Audebert, "Effect of water stress on growth, yield and yield components of rice (Oryza sativa L.) genotypes," Int. J. Sci. Res. Arch., vol. 5, no. 1, pp. 028–038, 2022, doi: 10.30574/ijsra.2022.5.1.0030.
18. S. S. Berutu, H. Budiati, Jatmika, F. Gulo, "Data Preprocessing Approach for Machine Learning-Based Sentiment Classification," *Jurnal Infotel*, Vol. 15, No. 4, h. 317-325, 2023.
19. I. O. Muraina, "Ideal Dataset Splitting Ratios in Machine Learning Algorithms: General Concerns for Data Scientists and Data Analysts," *7th International Mardin Artuklu Scientific Researches Conference*, h. 496-504, 2022.
20. B. T. Jijo., A. M. Abdulazeez, " Classification Based on Decision Tree Algorithm for Machine Learning," Vol. 2, No. 1, h. 20-28, 2021.
21. Z. Jun, "The Development and Application of Support Vector Machine," *Journal of Physics: Conference Series*, h. 1-8, 2020.
22. J. Cervantes, F. Garcia-Lamont, L. Rodriguez-Mazahua, A. Lopez, "A Comprehensive Survey on Support Vector Machine Classification: Applications, Challenges, and Trends," *Neurocomputing*, Vol. 408, h. 189-215, 2020.
23. J. Xu, Y. Zhang, D. Miao, "Three-Way Confusion Matrix for Classification: A Measure Driven View," *Information Sciences*, Vol. 507, h. 772-794, 2020.
24. A. Ali, A. Khairan, F. Tempola, and A. Fuad, "Application Of Naïve Bayes to Predict the Potential of Rain in Ternate City," *E3S Web Conf.*, vol. 328, p. 04011, 2021, doi: 10.1051/e3sconf/202132804011.