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

Perbandingan Algoritma *Machine Learning* dalam Klasifikasi Status Banjir di Sumatera Utara

Fatimah Rahmasari[1], Marsha Rifany[2], Teguh Priharyanto[3], Robert Kurniawan[4]
[1, 2, 3]*Program Studi Statistika, Politeknik Statistika STIS*
[4]*Program Studi Komputasi Statistik, Politeknik Statistika STIS*
[1]212112051@stis.ac.id
[2]212112178@stis.ac.id
[3]212112395@stis.ac.id
[4]robertk@stis.ac.id
Corresponding author email: 212112051@stis.ac.id

Abstract: *Indonesia is a country that is prone to experiencing hydrometeorological disasters, one of which is flooding. According to BNPB, floods are the most frequent disaster in Indonesia after earthquakes and landslides. In 2023, the region that dominates flooding in Indonesia will be Sumatra Island with North Sumatra Province as the province with the highest number of flood cases at 112 incidents in one year. Therefore, this research aims to find out the best model for classifying flood events in North Sumatra. The models compared are Decision Tree, Naive Bayes, KNN, Random Forest, and SVM. The parameters used are NDVI, rainfall, and slope. The results showed that the Random Forest model was the best model that could classify flood events into floods and non-floods with an accuracy of 95.59%.*
Keywords: *Flood, machine learning, classification, random forest*
Abstrak: Indonesia merupakan negara yang rawan mengalami bencana hidrometeorologi, salah satunya adalah banjir. Menurut BNPB, banjir merupakan bencana yang paling sering terjadi di Indonesia setelah gempa bumi dan tanah longsor. Pada 2023, wilayah yang paling mendominasi banjir di Indonesia adalah Pulau Sumatera dengan Provinsi Sumatera Utara sebagai provinsi dengan jumlah kasus banjir tertinggi sebesar 112 kejadian dalam kurun waktu satu tahun. Oleh karena itu, penelitian ini bertujuan untuk mengetahui model terbaik dalam klasifikasi kejadian banjir di Sumatera Utara. Model yang dibandingkan adalah *Decision Tree, Naive Bayes*, KNN, *Random Forest*, dan SVM. Parameter yang digunakan adalah NDVI, curah hujan, dan *slope*. Didapatkan hasil bahwa model *Random Forest* merupakan model terbaik yang dapat mengklasifikasikan kejadian banjir menjadi banjir dan tidak banjir dengan akurasi sebesar 95,59%.
Kata kunci: Banjir, machine learning, klasifikasi, random forest

I. PENDAHULUAN

Bencana merupakan kejadian atau peristiwa yang dapat memberikan kerugian yang besar dan bersifat merusak. Hal ini sering kali bersifat merugikan dan mengambil waktu yang panjang untuk melakukan pemulihannya. Salah satu bencana di Indonesia disebabkan oleh adanya curah hujan yang tinggi di Indonesia. Kondisi ini mengakibatkan Indonesia rawan mengalami bencana hidrometeorologi, salah satunya adalah banjir. Banjir adalah suatu peristiwa alam dimana terjadi luapan air yang mengakibatkan terbentuknya area tergenang. Bencana ini sering menyebabkan kerusakan ekonomi, sosial, fisik, dan budaya [1]. Menurut data dari Badan Nasional Penanggulangan Bencana (BNPB), pada tahun 2023 bencana banjir termasuk bencana yang paling sering terjadi di Indonesia setelah bencana gempa bumi dan tanah longsor. Jumlah kejadian banjir pada tahun 2023 mencapai angka 348 untuk seluruh wilayah di Indonesia.

Dari banyaknya kasus bencana banjir di Indonesia, wilayah yang paling mendominasi angka tersebut adalah berada pada pulau Sumatera dengan 233 kasus. Di antara beberapa provinsi yang berada pada pulau tersebut, Sumatera Utara mengalami bencana banjir yang paling banyak, yaitu sebanyak 112 kejadian banjir pada tahun 2023. Tentunya angka ini merupakan angka yang sangat tinggi untuk ukuran kejadian bencana yang dialami suatu wilayah dalam kurun waktu satu tahun. Bencana banjir di wilayah

tersebut mengakibatkan berbagai macam korban. Menurut data dari BNPB, adanya bencana banjir di Provinsi Sumatera Utara mengakibatkan 10 orang korban meninggal, 11 orang hilang, 102 orang terluka, dan memberi efek penderitaan kepada sebanyak 158 ribu orang di provinsi tersebut.

Banyaknya bencana banjir pada Provinsi Sumatera Utara memerlukan persiapan pra bencana untuk mengetahui mengenai faktor yang dapat mempengaruhi kejadian banjir pada suatu daerah secara spesifik melalui metode pengklasifikasian. Metode ini sangat berguna untuk dapat melakukan pemetaan wilayah berisiko banjir, perencanaan untuk melakukan mitigasi bencana banjir, pengelolaan sumber daya infrastruktur dan bantuan, serta untuk melakukan identifikasi daerah berpotensi banjir. Pengklasifikasian ini akan dibuat berdasarkan data yang dikumpulkan untuk mengenali daerah yang berpotensi mengalami banjir melalui faktor dependennya seperti daerah tutupan vegetasi, curah hujan, serta kemiringan lahan.

Sudah terdapat beberapa penelitian yang meneliti tentang banjir sebelumnya. Suwarsono [2] meneliti untuk melakukan identifikasi terhadap daerah wilayah banjir pada dataran rendah Purworejo, Bengawan Solo, dan wilayah hilir di Citarum. Ia membandingkan beberapa parameter NDWI seperti NDWI of Gao (1996), NDWI of McFeetrers (1996), MNDWI of Xu (2006), dan NWI of Yang (2011) untuk mengetahui parameter mana yang menghasilkan visualisasi daerah banjir yang lebih bagus. Wang dkk [3] meneliti mengenai pembentukan risiko bahaya banjir menggunakan model *random forest*. Variabel yang digunakan adalah curah hujan, frekuensi topan, model elevasi digital, indeks kelembaban topografi, NDVI, indeks daya aliran, tekstur tanah, jarak dengan sungai, kemiringan, dan kedalaman limpasan. Selain itu, penelitian Motta [4] juga dilakukan untuk memprediksi banjir menggunakan *machine learning* dan GIS. Variabel yang digunakan adalah temperatur, kelembaban, curah hujan, paparan sinar matahari, dan angin. Ia membandingkan beberapa metode klasifikasi seperti regresi logistik, SVM, Gaussian Naive-Bayes, *random forest*, *k-nearest neighbors*, dan *multi-layer perceptron* untuk mendapatkan model yang terbaik. Perbandingan model klasifikasi untuk banjir di Indonesia telah dilakukan oleh Priscillia dkk [5] dengan sampel 260 desa di Jakarta. Namun, pada penelitian sebelumnya yang telah disebutkan, belum ada yang memilih lokus Sumatera Utara dengan tingkat analisis kabupaten dan kota. Padahal, menurut BNPB [6], Sumatera Utara merupakan provinsi di Indonesia yang paling sering terkena banjir dan menyebabkan dampak yang cukup merugikan pada tahun 2023.

Berdasarkan penjelasan di atas, maka tujuan penelitian ini adalah untuk mengetahui hasil klasifikasi wilayah potensi banjir dari beberapa metode pengklasifikasian dengan *machine learning* dan memilih model dengan akurasi terbaik berdasarkan variabel yang memengaruhinya. Beberapa model yang akan digunakan untuk dapat menjawab tujuan ini adalah *decision tree*, Naïve Bayes, *k-Nearest Neighbors* (KNN), *random forest*, dan *Support Vector Machine* (SVM). Hasil penelitian ini diharapkan dapat mengidentifikasi model terbaik untuk prediksi kejadian banjir yang dapat digunakan sebagai upaya mitigasi risiko banjir di Sumatera Utara.

II. METODE PENELITIAN

Pada penelitian ini, tahapan yang dilakukan mencakup empat sub tahapan, yaitu *Data Understanding, Data Preparation, Modelling,* dan *Evaluation*.

II.1. Data Understanding

Pada tahap ini, dilakukan eksplorasi awal terhadap dataset kejadian banjir sebagai variabel target dan beberapa variabel parameter. Unit observasi pada penelitian ini mencakup 455 kecamatan di Sumatera Utara. Dataset didapatkan dari sumber primer, yaitu data citra satelit dan diambil melalui

Google Earth Engine. Variabel parameter yang digunakan adalah NDVI, curah hujan, dan kemiringan lereng (*slope*). Data ini diambil untuk periode waktu 1 November 2023 hingga 30 November 2023. Sementara itu, data kejadian banjir didapatkan dari nilai ΔNDWI antara periode 1 Januari 2022-31 Januari 2022 dan periode 1 November 2023-30 November 2023. Periode di bulan Januari 2022 dipilih karena berdasarkan data dari BNPB, hanya terdapat 1 kejadian banjir di Sumatera Utara. Periode di bulan November 2023 dipilih karena pada periode tersebut, terdapat jumlah kejadian banjir tertinggi di Sumatera Utara. Nilai NDWI yang lebih besar pada November 2023 mengindikasikan terdapat kejadian banjir. Lalu, berdasarkan data NDWI November 2023 dan *change* NDWI, didapatkan variabel status kejadian banjir dengan dua kategori sebagai berikut:

- 0, jika tidak terjadi banjir pada kecamatan tertentu;
- 1, jika terjadi banjir pada kecamatan tertentu.

Suatu kecamatan dikatakan banjir apabila memiliki ΔNDWI ≥ 0,094 dan $NDWI_{Nov\ 2023} \geq$ 0,161 [2]. Berikut ini adalah deskripsi dari variabel yang digunakan.

Tabel 1. Variabel yang Digunakan

Variabel	Tipe Data	Satuan
NDVI	Kontinu	-
Curah Hujan	Kontinu	mm
Slope	Kontinu	Derajat
Status Banjir	Kategorik Biner	-

II.2. Data Understanding

Pada tahap ini, dilakukan persiapan data hingga didapatkan data yang siap untuk dianalisis. Hal ini mencakup pengecekan *extreme value* dan *missing value*. *Extreme value* yang terdapat pada dataset dihapus. Sementara itu, jika terdapat *missing value* maka akan dilakukan imputasi dengan menggunakan rata-rata.

II.3. Modeling

Data yang digunakan akan dibagi menjadi 2 subset dengan rasio 70:30, di mana 70% data digunakan untuk *training* dan 30% digunakan untuk *testing*. Pada penelitian ini, dilakukan pemodelan dengan beberapa model klasifikasi, seperti *decision tree*, Naïve Bayes, *k-Nearest Neighbors* (KNN), *random forest*, dan *Support Vector Machine* (SVM).

II.3.1 Decision Tree

Decision Tree merupakan metode pembelajaran mesin yang digunakan untuk klasifikasi dengan memodelkan keputusan yang mengarah pada hasil tertentu berdasarkan fitur *input*. Proses pembentukan *decision tree* melibatkan pembagian data ke dalam subset yang lebih kecil dengan homogenitas yang semakin meningkat melalui pemilihan fitur yang paling informatif pada setiap *node*, berdasarkan kriteria seperti Information Gain atau Gini Index [7].

II.3.2 Naïve Bayes

Klasifikasi Naïve Bayes dibangun berdasarkan teorema Bayes, dengan asumsi bahwa ada independensi antara prediktor-prediktor. Model Naïve Bayesian sangat mudah dibangun dan sangat berguna untuk dataset berukuran besar karena tidak memerlukan estimasi parameter yang kompleks secara iteratif. Kinerja klasifikasi Naïve Bayes sering kali sangat baik meskipun modelnya cukup sederhana, bahkan sering kali mengungguli metode klasifikasi yang lebih kompleks, sehingga banyak digunakan. Dengan menggunakan teorema Bayes, kita dapat menentukan probabilitas

terjadinya suatu peristiwa, dengan asumsi kita memiliki probabilitas peristiwa lain yang sudah terjadi [7].

II.3.3. K-Nearest Neighbors

KNN adalah algoritma non parametrik yang digunakan untuk klasifikasi dan regresi. *Input* terdiri dari titik-titik terdekat yang signifikan dan *output* bergantung pada klasifikasi atau regresi algoritma KNN. Klasifikasi dilakukan berdasarkan *majority votes*. Objek ditempatkan ke dalam kelas yang paling sering muncul di antara k tetangga terdekatnya. Ketika nilai k sama dengan satu, objek ditempatkan ke tetangga terdekatnya. Untuk regresi KNN, nilai properti yang ditemukan menjadi output.

Masalah yang sering terjadi pada algoritma KNN adalah menentukan nilai k. Semakin kecil nilai k berakibat pada *noise* yang berpengaruh makin besar terhadap *output*, misalnya menyebabkan kemungkinan *overfitting*. Untuk itu, penentuan nilai k dapat dilakukan dengan menggunakan *Cross Validation*. Model yang memberikan akurasi terbaik dianggap sebagai pilihan nilai k terbaik [8].

II.3.4. Random Forest

Random Forest adalah algoritma yang digunakan untuk klasifikasi dengan membangun sejumlah pohon keputusan (*decision trees*) selama pelatihan dan menggabungkan prediksi dari masing-masing pohon untuk menghasilkan prediksi akhir yang lebih akurat dan *robust* [9]. Setiap pohon dalam *forest* dibangun dari sampel *bootstrap* dari data *training*, dan setiap pemisahan *node* dalam pohon dipilih dari subset acak dari fitur, yang mengurangi varian dan risiko *overfitting*. Model prediksi akhir dihasilkan dengan menggabungkan prediksi dari semua melalui metode voting mayoritas untuk klasifikasi [8].

II.3.5. Support Vector Machine (SVM)

SVM merupakan algoritma yang digunakan untuk menganalisis data dalam studi regresi dan klasifikasi. Algoritma SVM pada data training pada dasarnya menciptakan model yang menunjuk contoh ke satu kategori atau yang lain, mengembangkan pengklasifikasi linier biner non-probabilistik [10].

II.4. Evaluation

Dari beberapa model yang telah dibangun, selanjutnya dilakukan evaluasi model dengan beberapa ukuran untuk mengetahui model mana yang memiliki kemampuan terbaik untuk klasifikasi kejadian banjir. Beberapa ukuran yang digunakan adalah sebagai berikut.

II.4.1 Confusion matrix

Karena variabel target yang digunakan memiliki 2 kategori, maka *confusion matrix* yang terbentuk berukuran 2x2. Bentuk dari *confusion matrix* yang akan dibentuk disajikan pada tabel 2.

Tabel 2. Confusion Matrix

		Prediksi	
		Positive	*Negative*
Aktual	*Positive*	*True Positive* (TP)	*False Negative* (FN)
	Negative	*False Positive* (FP)	*True Negative* (TN)

II.4.2. Accuracy

Akurasi adalah ukuran umum yang digunakan untuk mengevaluasi kinerja model klasifikasi. Ini mengukur seberapa sering model benar-benar memprediksi kelas dengan benar, baik

positif maupun negatif. Semakin tinggi nilai akurasi, semakin baik model dalam membuat prediksi yang benar. Namun, akurasi tidak selalu merupakan metrik yang tepat untuk dievaluasi jika dataset memiliki kelas yang tidak seimbang (*imbalance*), karena dapat menyesatkan dalam kasus tersebut. Oleh karena itu, penting untuk mempertimbangkan metrik evaluasi lainnya seperti precision, recall, dan F1 Score bersama dengan akurasi. Rumus hitung dari akurasi adalah sebagai berikut.

$$Akurasi = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

[11]

II.4.3 Precision

Precision memberikan proporsi prediksi positif yang benar terhadap semua prediksi positif. *Precision* yang tinggi berarti sebagian besar prediksi positif model adalah benar. Rumus hitung dari *precision* adalah sebagai berikut.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

II.4.4. Recall (Sensitivity)

Recall mengukur seberapa banyak dari total kasus positif yang berhasil terdeteksi oleh model. Dengan kata lain, *recall* adalah proporsi kasus positif aktual yang berhasil diprediksi dengan benar oleh model. Rumus hitung dari *recall* adalah sebagai berikut.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

II.4.5. F1 Score

F1 Score merupakan rata-rata harmonis antara *precision* dan *recall*. *F1 Score* mencapai nilai terbaik di 1 dan nilai terburuk di 0. Rumus hitung dari *F1 Score* adalah sebagai berikut.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

[12]

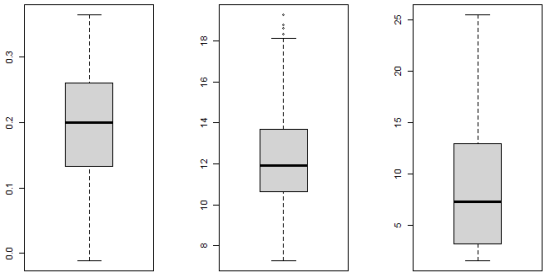
III. HASIL DAN PEMBAHASAN

Pada penelitian ini, tahapan yang dilakukan mencakup empat sub tahapan, yaitu *Data Understanding, Data Preparation, Modelling,* dan *Evaluation*.

III.1. Data Preprocessing

Setelah dilakukan tahap *data preprocessing,* ditemukan adanya *missing value* pada variabel *Slope* sebanyak 1 amatan sehingga penulis memutuskan untuk mengimputasi. Imputasi yang digunakan adalah nilai *mean*. Setelah imputasi *missing value*, dilakukan visualisasi data menggunakan *boxplot* untuk mendeteksi adanya *outlier*.

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



Gambar 1. Boxplot NDVI, Curah Hujan, dan Slope

Berdasarkan gambar 1, terlihat bahwa dari ketiga variabel, curah hujan memiliki *outlier* atas. Namun, data tersebut tetap dipertahankan di dalam dataset untuk kemudian dianalisis.

II.2. Pemodelan

Sebelum melakukan pemodelan, dilakukan deteksi multikolinearitas terhadap seluruh variabel yang digunakan. Berdasarkan tabel 3, diketahui bahwa seluruh variabel memiliki nilai VIF < 5 yang mengindikasikan tidak terjadi multikolinearitas. Karena tidak terdapat multikolinearitas, maka pemodelan dapat dilakukan.

Tabel 3. VIF Parameter

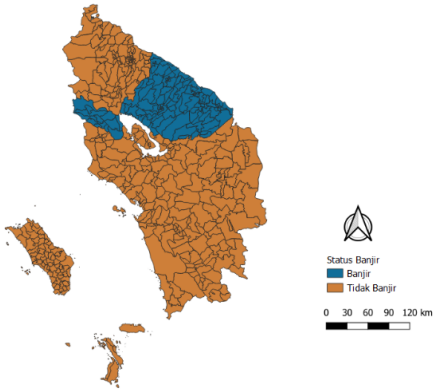
Variabel	VIF
NDVI	1,0591
Curah Hujan	1,3848
Slope	1,0261

Sebelum dilakukan pemodelan, berikut ini disajikan hasil analisis deskriptif melalui visualisasi dan tabel ringkasan statistik. Visualisasi NDWI di Sumatera Utara pada bulan Januari 2022 (waktu kering) dan November 2022 (waktu banjir) dapat dilihat pada gambar 2.

Gambar 2. Visualisasi NDWI (2a) Visualisasi NDWI Januari 2022 (2b) Visualisasi NDWI November 2023

Berdasarkan gambar 2, terlihat bahwa nilai NDWI pada November 2023 lebih besar dibandingkan Januari 2022 yang ditunjukkan oleh area berwarna biru gelap yang lebih luas. Hal ini menunjukkan bahwa area yang mengandung perairan di permukaannya semakin luas dibandingkan periode waktu sebelumnya. Kondisi ini mengindikasikan bahwa terdapat genangan air/banjir yang terjadi pada wilayah tersebut. Berdasarkan nilai ΔNDWI dan *threshold* yang digunakan, didapatkan status kejadian banjir untuk setiap kecamatan di Sumatera Utara dan hasil visualisasinya terdapat pada gambar berikut.

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



Gambar 3. Status Banjir Kecamatan di Sumatera Utara November 2023

Berdasarkan gambar 3, terlihat bahwa kecamatan yang mengalami banjir cenderung mengelompok di Sumatera Utara bagian tengah.

Karakteristik dari wilayah yang mengalami banjir dan tidak banjir disajikan pada tabel ringkasan statistik berikut.

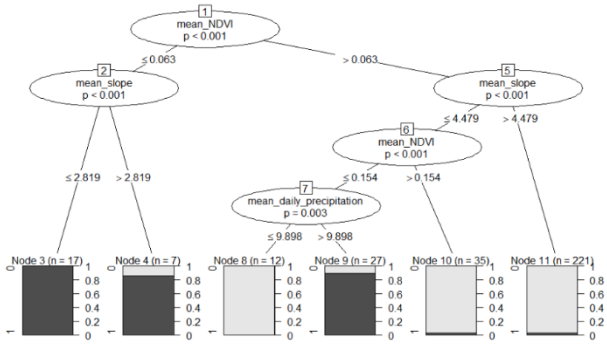
Tabel 4. Rata-rata Parameter

Status Banjir	Rata-rata NDVI	Rata-rata Curah Hujan	Rata-rata *Slope*
Banjir	0,07589	11,04385	3,67892
Tidak Banjir	0,21689	12,64338	9,52095

Berdasarkan tabel 4, diketahui bahwa wilayah yang banjir memiliki rata-rata NDVI, curah hujan, dan *slope* yang lebih kecil dibandingkan wilayah yang tidak banjir. Setelah itu dilakukan pemodelan menggunakan lima model yang berbeda antara lain: *Decision Tree*, Naive Bayes, *k-Nearest Neighbors*, *Random Forest* dan *Support Vector Machine*.

III.2.1. Decision Tree

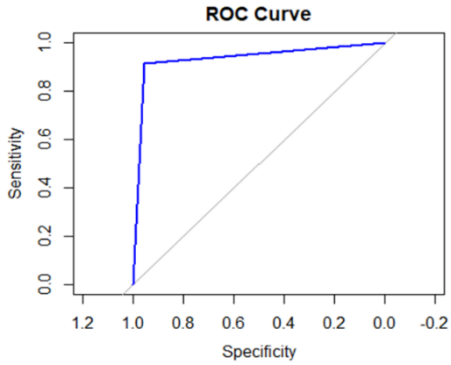
Hasil pemodelan menggunakan *decision tree* menghasilkan pohon sebagai berikut.



Gambar 4. Pohon Pemodelan Decision Tree

Berdasarkan hasil model *decision tree* pada gambar 4, *root node* adalah mean_NDVI. ketika mean_NDVI bernilai <0,063 dan mean_slope <2,819, maka kemungkinan terjadi banjir sangat tinggi. namun jika mean_slope >2,819 kemungkinan tersebut sedikit berkurang walaupun masih tergolong tinggi. Namun jika mean_NDVI bernilai >0,063 dan mean_slope >4,479 maka kemungkinan terjadi banjir sangat rendah. begitu pula saat mean_slope <4,479 namun dengan

syarat mean_NDWI >0,154. Saat mean_NDWI >0,154 dan mean_daily_precipitation <9,898 memiliki kemungkinan banjir yang sangat rendah berbeda saat mean_daily_precipitation >9,898 memiliki kemungkinan banjir yang tinggi.

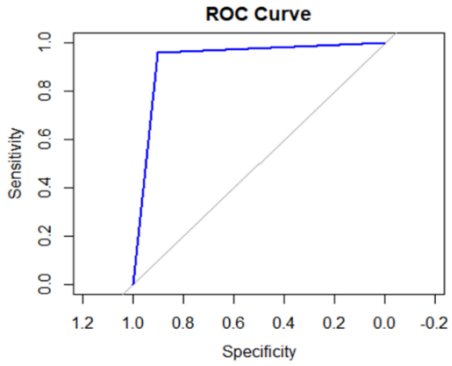


Gambar 5. Kurva ROC Model Decision Tree

Kurva ROC pada gambar 5 menunjukkan garis yang memiliki sudut mendekati sudut kiri atas, yang artinya semakin baik kinerja model *decision tree*. Nilai AUC dari model ini juga sebesar 0,9343978, angka yang mendekati nilai maksimum yaitu 1 menunjukkan kinerja model tersebut baik.

III.2.2 Naive Bayes

Pemodelan menggunakan Naive Bayes menghasilkan kurva ROC sebagai berikut.



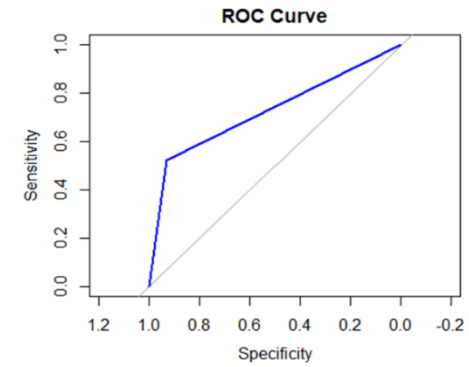
Gambar 6. Kurva ROC Model Naive Bayes

Berdasarkan kurva AUC pada gambar 6, ditunjukkan garis yang memiliki sudut mendekati sudut kiri atas walaupun dibandingkan dengan model sebelumnya memiliki nilai *sensitivity* yang lebih tinggi namun memiliki *specificity* yang lebih rendah, yang artinya semakin baik kinerja model Naive Bayes. Nilai AUC dari model ini juga sebesar 0,9295883, angka yang mendekati nilai maksimum yaitu 1 menunjukkan kinerja model tersebut baik.

III.2.3. K-Nearest Neighbors

Dalam pemodelan, dilakukan pencarian parameter yang optimum supaya menghasilkan model yang memiliki performa terbaik untuk klasifikasi. Dalam model KNN, perlu ditentukan nilai k yang optimal berdasarkan nilai *Cross Validation*. Didapatkan bahwa nilai Cross Validation terbesar dicapai ketika k bernilai 7. Oleh karena itu, pemodelan KNN dilakukan dengan

menggunakan k sebesar 7. Selanjutnya dilakukan evaluasi model menggunakan visualisasi melalui kurva ROC.



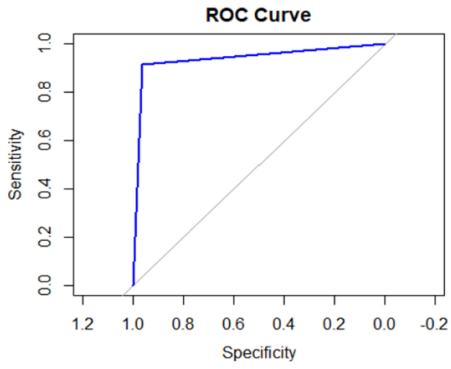
.

Gambar 7. Kurva ROC Model KNN

Berdasarkan kurva ROC pada gambar 7, ditunjukkan garis yang memiliki sudut memiliki arah ke sudut kiri atas walaupun masih terlalu jauh untuk dikatakan mendekati sudut kiri atas dibandingkan dengan model sebelumnya, yang artinya baik kinerja model KNN. Nilai AUC dari model ini juga sebesar 0,7254713, angka yang mengarah ke nilai maksimum yaitu 1 menunjukkan kinerja model tersebut cukup baik.

III.2.4. Random Forest

Pencarian parameter yang optimum pada *model random forest* dilakukan pada penelitian ini. Berikut merupakan hasil pencarian parameter optimum yang disajikan melalui grafik di bawah. Diketahui bahwa parameter optimum yang diberikan adalah ntree = 600 dan mtry = 5.



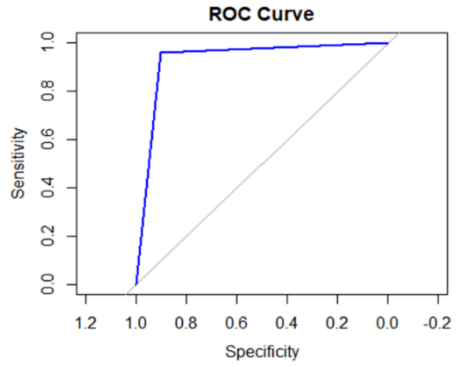
Gambar 8. Kurva ROC Model Random Forest

Berdasarkan kurva ROC pada gambar 8, ditunjukkan garis yang memiliki sudut mendekati arah ke sudut kiri atas dan lebih baik dibandingkan dengan model sebelumnya, yang artinya sangat baik kinerja model *random forest*. Nilai AUC dari model ini juga sebesar 0,9388226, angka yang mendekati ke nilai maksimum yaitu 1 menunjukkan kinerja model tersebut sangat baik.

III.2.5. Support Vector Machine

Pemodelan menggunakan SVM menghasilkan kurva ROC sebagai berikut.

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



Gambar 9. Kurva ROC Model Naive Bayes

Berdasarkan kurva ROC pada gambar 9, ditunjukkan garis yang memiliki sudut mendekati arah ke sudut kiri atas walaupun memiliki *sensitivity* lebih rendah dibandingkan dengan model sebelumnya, yang artinya sangat baik kinerja model SVM. Nilai AUC dari model ini juga sebesar 0,9126587, angka yang mendekati ke nilai maksimum yaitu 1 menunjukkan kinerja model tersebut sangat baik.

Untuk mengetahui model terbaik, dilakukan evaluasi menggunakan beberapa ukuran seperti yang disajikan pada tabel 5 di bawah ini.

Tabel 5. Evaluasi Model

Ukuran Evaluasi	Decision Tree	Naive Bayes	KNN	Random Forest	SVM
Accuracy	0,9485	0,9118	0,8603	**0,9559**	0,9412
Precision	0,913	0,9565	0,5217	**0,913**	0,8696
Recall	0,8077	0,6667	0,6	**0,84**	0,8
F1-Score	0,8571	0,6774	0,5581	**0,875**	0,8333
AUC Score	0,9344	0,9296	0,7255	**0,9388**	0,9127

Berdasarkan tabel 5, didapatkan hasil bahwa model klasifikasi dengan algoritma *Random Forest* memiliki nilai *Accuracy, Precision, Recall* dan *F1 Score* berturut-turut sebesar 95,59%; 91,30%; 84%; dan 87,5%. Nilai ini paling besar dibandingkan keempat model lainnya yang menunjukkan bahwa model *random forest* merupakan model terbaik.

Berdasarkan hasil penelitian di atas, diketahui bahwa model klasifikasi terbaik untuk status kejadian banjir di Sumatera Utara adalah *Random Forest*. Hal ini sejalan dengan penelitian oleh Seydi dkk [13] yang menyatakan bahwa model dengan algoritma *Random Forest* lebih baik dibandingkan SVM dan *decision tree* dalam mengklasifikasikan kerawanan banjir di cekungan Gorganrud dan Karun di Iran. Hasanuzzaman dkk [14] dalam penelitiannya juga mendapatkan hasil bahwa model Random Forest memiliki performa lebih baik dibandingkan model Naive Bayes dan *Extreme Gradient Boosting* dalam mengklasifikasikan kerawanan banjir di sungai Silabati, India. Hasil penelitian oleh Dominic dan Kurian [15] juga menyatakan bahwa model Random Forest lebih baik dibandingkan KNN dan *decision tree* dalam memprediksi banjir. Penelitian lain oleh Chen dkk [16] juga menunjukkan hasil bahwa antara pemodelan *random forest*, Naive Bayes dan *decision tree*, ketiganya menunjukkan kurva ROC yang baik, namun *random forest* memiliki nilai PPR, NPR dan akurasi yang lebih tinggi dibandingkan Naive Bayes dan *decision tree*. Hasil akhir *random*

forest mengungguli model lainnya untuk kerentanan banjir dan telah dikonfirmasi penelitian lain [17][18].

Model *random forest* memiliki performa yang lebih baik dibandingkan *decision tree* dikarenakan *decision tree* merupakan algoritma yang sensitif terhadap data yang rentan terhadap overfitting sehingga mengurangi keakuratan pada dataset selain data pelatihan. Sementara itu, model RF telah meningkatkan keakuratannya setelah *ensemble*, karena *ensemble* dapat mengurangi pengaruh *overfitting* atau *underfitting* dari *decision tree* dasar [19]. Selain itu, *random forest* dapat efisien dengan database besar dan memberikan klasifikasi yang jelas hasil dengan menggabungkan banyak *decision tree* [20].

IV. KESIMPULAN

Didapatkan hasil bahwa wilayah yang banjir memiliki rata-rata NDWI, curah hujan, dan *slope* lebih kecil dibandingkan wilayah yang tidak banjir. Ketiga variabel tersebut juga sudah cukup baik untuk membuat klasifikasi status banjir di tiap kecamatan pada Provinsi Sumatera Utara. Model terbaik yang dapat mengklasifikasikan kejadian banjir di Sumatera Utara adalah *random forest* dengan akurasi sebesar 95,59%.

Saran untuk penelitian selanjutnya, untuk meningkatkan akurasi bisa menambahkan parameter yang digunakan dalam pengklasifikasian serta menggunakan metode klasifikasi yang lebih *advance*. Hal ini dapat dilakukan untuk mendapatkan hasil dengan akurasi yang lebih tinggi. Selain itu juga memperhatikan kondisi data *imbalance* untuk menghindari terjadinya kemungkinan *overfitting*.

REFERENSI

1. C. J. Talbot *et al.*, "The impact of flooding on aquatic ecosystem services," *Biogeochemistry*, vol. 141, no. 3, pp. 439–461, 2018, doi: 10.1007/s10533-018-0449-7.
2. Suwarsono, N. J. Tejo, and Wiweka, "Identification of Inundated Area Using Normalized Difference," *Int. J. Remote Sens. Earth Sci.*, vol. 10, no. 2, pp. 114–121, 2013.
3. Z. Wang, C. Lai, X. Chen, B. Yang, S. Zhao, and X. Bai, "Flood hazard risk assessment model based on random forest," *J. Hydrol.*, vol. 527, pp. 1130–1141, 2015, doi: 10.1016/j.jhydrol.2015.06.008.
4. M. Motta, M. de Castro Neto, and P. Sarmento, "A mixed approach for urban flood prediction using Machine Learning and GIS," *Int. J. Disaster Risk Reduct.*, vol. 56, no. February, p. 102154, 2021, doi: 10.1016/j.ijdrr.2021.102154.
5. S. Priscillia, C. Shilacci, and A. Lipani, "Flood usceptibility assessment using artificial neural networks in Indonesia," *Artif. Intell. Geosci.*, vol. 2, pp. 215–222, 2021, doi: https://doi.org/10.1016/j.aiig.2022.03.002.
6. B. N. P. Bencana, "Data Informasi Bencana Indonesia," 2023. https://dibi.bnpb.go.id/kwilayah2
7. R. Genuer and J.-M. Poggi, *Introduction to Random Forests with R*. 2020. doi: 10.1007/978-3-030-56485-8_1.
8. A. Gereon, *Hands-on Machine Learning with Scikit Learn*, Keras & Te. O'Reilly Media.
9. J. M. Sadler, J. L. Goodall, M. M. Morsy, and K. Spencer, "Modeling urban coastal flood severity from crowd-sourced flood reports using Poisson regression and Random Forest," *J. Hydrol.*, vol. 559, pp. 43–55, 2018, doi: 10.1016/j.jhydrol.2018.01.044.
10. S. Suthaharan, *Machine Learning Models and Algorithms for Big Data Classifiocation*. Springer, 2016.
11. A. Zheng, *Evaluating Machine Learning Models*, First Edit. Sebastopol, CA: O'Reilly media, 2021. doi: 10.1007/978-1-4842-6537-6_7.
12. H. Dalianis, "Evaluation Metrics and Evaluation," *Clin. Text Min.*, no. 1967, pp. 45–53, 2018, doi: 10.1007/978-3-319-78503-5_6.
13. S. T. Seydi, Y. Kanani-Sadat, M. Hasanlou, R. Sahraei, J. Chanussot, and M. Amani, "Comparison of Machine Learning Algorithms for Flood Susceptibility Mapping," *Remote Sens.*, vol. 15, no. 1, 2023, doi: 10.3390/rs15010192.
14. M. Hasanuzzaman, A. Islam, B. Bera, and P. K. Shit, "A comparison of performance measures of three

machine learning algorithms for flood susceptibility mapping of river Silabati (tropical river, India)," *Phys. Chem. Earth, Parts A/B/C*, vol. 127, 2022, doi: https://doi.org/10.1016/j.pce.2022.103198.
15. D. Kiran and R. Kurian, "Comparative Study of Machine Learning Algorithms for Intrusion Detection," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 4s, pp. 647–653, 2024, doi: 10.5281/zenodo.6369640.
16. W. Chen *et al.*, "Modeling flood susceptibility using data-driven approaches of naïve Bayes tree, alternating decision tree, and random forest methods," *Sci. Total Environ.*, vol. 701, p. 134979, 2019, doi: 10.1016/j.scitotenv.2019.134979.
17. B. Choubin, E. Moradi, M. Golshan, J. Adamowski, F. Sajedi-Hosseini, and A. Mosavi, "An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines," *Sci. Total Environ.*, vol. 651, pp. 2087–2096, 2019, doi: 10.1016/j.scitotenv.2018.10.064.
18. G. Zhao, B. Pang, Z. Xu, D. Peng, and L. Xu, "Assessment of urban flood susceptibility using semi-supervised machine learning model," *Sci. Total Environ.*, vol. 659, pp. 940–949, 2019, doi: 10.1016/j.scitotenv.2018.12.217.
19. R. Wang, "Comparison of Decision Tree, Random Forest and Linear Discriminant Analysis Models in Breast Cancer Prediction," *J. Phys. Conf. Ser.*, vol. 2386, no. 1, 2022, doi: 10.1088/1742-6596/2386/1/012043.
20. N. Khan, S. Shahid, L. Juneng, K. Ahmed, T. Ismail, and N. Nawaz, "Prediction of heat waves in Pakistan using quantile regression forests," *Atmos. Res.*, vol. 221, no. November 2018, pp. 1–11, 2019, doi: 10.1016/j.atmosres.2019.01.024.