e-ISBN 978-623-93343-4-5
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 15 September 2021

Perbandingan Metode Klasifikasi *Random Forest* dan *XGBoost* Serta Implementasi Teknik *SMOTE* pada Kasus Prediksi Hujan

Ghaitsa Amany Mursianto[1], Isma'il Muhammad Falih[2], Muhammad Irfan[3], Tiara Sakinah[4], Desta Sandya Prasvita[5]
Program Studi Informatika / Fakultas Ilmu Komputer
Universitas Pembangunan Nasional Veteran Jakarta
Jl. RS. Fatmawati Raya, Pd. Labu, Kec. Cilandak, Kota Depok, Daerah Khusus Ibukota Jakarta 12450
ghaitsaam@upnvj.ac.id[1], isma'ilmf@upnvj.ac.id[2], m.irfan@upnvj.ac.id[3],tiarasakinah@upnvj.ac.id[4], desta.sandya@upnvj.ac.id[5]

Abstrak. Cuaca pada hakikat merupakan suatu keadaan udara pada waktu tertentu di wilayah yang berbeda-beda dengan cakupan wilayah yang luas maupun sempit dalam jangka waktu yang relatif singkat. Cuaca juga merupakan suatu kombinasi dari beberapa komponen yaitu tekanan suhu, kecepatan angin dan arah mata angin, jumlah volume air yang terkandung pada awan, tekanan udara, kelembaban udara, dan lain-lain. Cuaca sendiri merupakan gambaran dari suatu fenomena alam yang tidak dapat diprediksi seperti hujan dengan intensitas padasaat musim kemarau, udara panas pada saat musim hujan, badai, dan lain-lain. Oleh sebab itu penelitian ini bertujuan untuk mengklasifikasi prediksi hujan pada hari-hari berikutnya, dengan menggunakan beberapa metode klasifikasi yaitu Random Forest, XGBoost dari metode tersebut akan menentukan kelas mana yang paling optimal. Sistem prediksi cuaca yang telah kami buat mendapatkan tingkat akurasi tertinggi diperoleh klasifikasi Random Forest dengan Resampling yakni sebesar 95.59%, namun pada klasifikasi tanpa resampling akurasi tertinggi diperoleh XGBoost yakni sebesar 94.34%.

Kata Kunci: Klasifikasi, Random Forest, XGBoost, Cuaca

1. Pendahuluan

Cuaca merupakan kombinasi dari beberapa komponen antara lain tekanan suhu, jumlah volume airyang terkandung pada awan, kecepatan angin dan arah mata angin, tekanan udara, kelembaban udara, dan lain-lain [1]. Prakiraan cuaca menjadi salah satu hal yang sangat dibutuhkan oleh orang-orang di seluruh dunia, agar tidak mengganggu kegiatan yang telah direncanakannya. Oleh karena itu diperlukan pengamatan yang lebih terhadap cuaca. Besarnya pengaruh iklim yang disebabkan, mengakibatkan pengembangan sistem cuaca untuk menentukan kondisi cuaca pada hari-hari berikutnya. Dengan adanya data historis mengenai keadaan cuaca yang terjadi di masa lampau, data tersebut dapat dipakai untuk menentukan cuaca yang akan terjadi di waktu yang akandatang. Penelitian terkait prediksi curah hujan sudah banyak dilakukan menggunakan berbagai metode, diantaranya metode random forest, C4.5, dan classification and regression trees (CART).

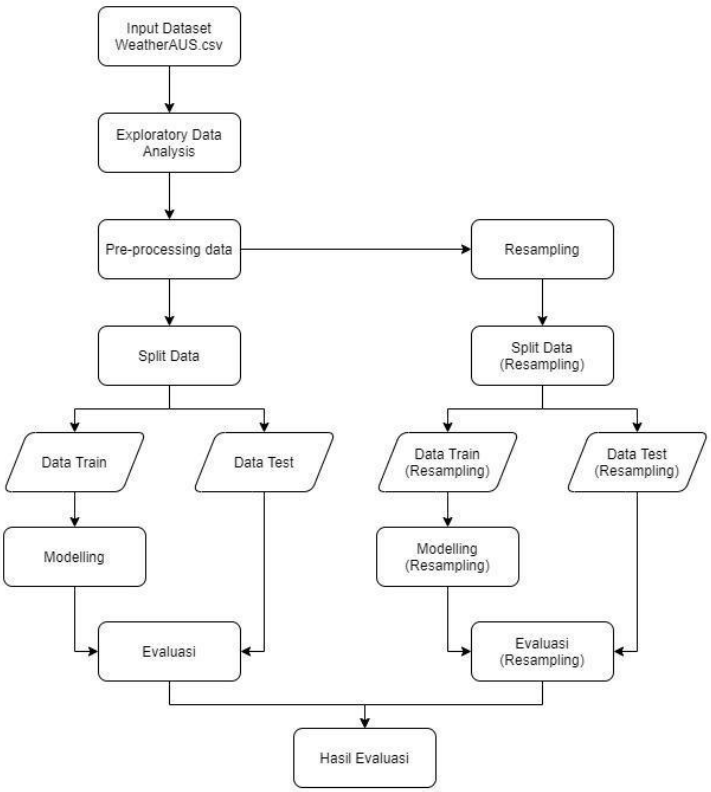
Pada penelitian ini menerapkan algoritma klasifikasi random forest dan XGBoost untuk memperkirakan curah hujan. Random forest melakukan penggabungan pohon/ tree dengan melakukan training pada data yang dimiliki [2]. Metode ini biasa digunakan karena menghasilkan kesalahan dengan persentase rendah, serta hasil akurasi yang didapat cukup tinggi dalam klasifikasi untuk jumlah data yang sangat besar.

Penelitian yang telah dilakukan sebelumnya yang menggunakan metode Random Forest pada citra seperti pada "PERBANDINGAN ALGORITMA KLASIFIKASI UNTUK PREDIKSI CUACA"
oleh Amril Mutoi Siregar, Sutan Faisal, Yana Cahyana, dan Bayu Priyatna yang mendapat akurasi sebesar 82.38% dengan standar deviasi sebesar 43% [3], dan penelitian lain yang menggunakan metode XGBoost pada citra seperti pada "Implementasi Metode XGBoost dan Feature Importance untuk Klasifikasi pada Kebakaran Hutan dan Lahan " oleh Ichwanul Muslim Karo Karo yang mendapat akurasi akurasi sebesar 89.52% [4].

Pada penelitian ini menggunakan kumpulan data bernama WeatherAUS dengan tipe csv yang diperoleh dari situs kaggle.com. Data WeatherAUS adalah data pengamatan cuaca harian dari berbagai lokasi di seluruh Australia, data ini diperoleh dari Biro Meteorologi Persemakmuran Australia. Data memiliki variabel target RainTomorrow

(apakah pada hari berikutnya akan hujan dengan indikator Tidak/Ya) dan variabel risiko RISK_MM (berapa banyak hujan yang tercatat dalam milimeter). Data yang diperoleh memiliki sebanyak 23 fitur dan sebanyak 145.460 data untuk dilakukan prediksi hujan. Kumpulan data tersebut berupa angka dan huruf yang nantinya akan diselaraskan menjadi angka.

2. Metodologi Penelitian



Gambar. 1. Tahapan Metode Penelitian

2.1 Input dataset

Data yang kami gunakan pada penelitian ini berasal dari Kaggle.com (https://www.kaggle.com/jsphyg/weather-dataset-rattle-package). Dataset tersebut berasal dari suatu perusahaan yang memiliki masalah tertentu. Dataset yang kami gunakan adalah kumpulan data cuaca dengan nama WeatherAUS berformat csv, data ini memiliki 30 fitur dengan jumlah data sebanyak 145.460 yang ditujukan untuk memprediksi hujan di keesokan hari dengan fitur RainTomorrow sebagai variabel target.

2.2 Pre-processing data

Data yang digunakan selanjutnya akan dilakukan pre-proses guna menghilangkan noise pada data WeatherAUS dan menyiapkan data untuk tahap selanjutnya. Noise dalam bahasa Indonesia adalah Derau atau data-data yang tidak diinginkan dan dapat mengganggu kualitas data. Pre proses terdiri dari beberapa tahapan, yakni Mengisi missing value, Mengubah tipe data, Menghapus fitur, Standarisasi, dan menghapus outlier.

2.2.1 Mengisi missing value

Salah satu noise yang terdapat pada data ini adalah missing value, maka dari itu dilakukan pengisian pada missing value menjadi modus atau nilai terbanyak. Missing value diisikan dengan nilai terbanyak dengan parameter .mode(). Setelah dilakukan pengisian missing value,dapat di cek kembali untuk melihat apakah masih ada missing value atau tidak.

2.2.2 Mengubah tipe data

Selanjutnya, fitur dataset yang masih memiliki tipe data objek, diubah menjadi tipe data integer, untuk memudahkan dalam melakukan modelling. Fitur - fitur yang diubah menjadi integer yaitu, WindGustDir, WindDir9AM, WindDir3PM, Rain today dan Rain tomorrow.

2.2.3 Menghapus Fitur

Kemudian dilakukan pemilihan fitur dengan cara menghapus fitur yang tidak mempengaruhi dalam proses klasifikasi dengan menggunakan fungsi .drop() untuk menghapus fitur, beberapa fitur yang dihapus yakni Date dan location.

2.2.4 Standarisasi

Standarisasi data merupakan salah satu proses yang penting dalam klasifikasi, standarisasi digunakan untuk merubah nilai dari variabel agar mudah dimengerti dan dibandingkan dengan variabel lainnya. Pada tahap ini semua variabel yang digunakan dilakukan standarisasi, mulai dari variabel MinTemp hingga variabel RainTomorrow.

2.2.5 Menghapus outlier

Setelah dilakukan standarisasi dilakukan visualisasi outlier terhadap semua variabel yang ada ,apakah masih terdapat outlier atau tidak untuk memudahkan dalam proses klasifikasinya, kemudian dilakukan penghapusan data variabel yang memiliki data outlier lebih dari 75 dan 25.

2.3 Resampling

Pada tahap ini dilakukan resampling dengan menggunakan SMOTE, SMOTE merupakan metode resampling untuk memperbanyak data pada kelas minoritas dengan cara menduplikasi data/ menggunakan data sintetik yang ada di kelas minoritas dengan tujuan mengurangi ketidakseimbangan data [5][6]. Dengan menggunakan SMOTE akan dapat menghindari masalah overfitting yang tinggi. Alasan digunakannya SMOTE karena data setelah praproses tergolong imbalance data moderate atau sedang dengan distribusi variabel target setelah praproses 89%: 11%.

2.4 Pembagian data

Data yang telah diperoleh akan dibedakan menjadi dua kelompok, yaitu data latih dan data uji. Penelitian ini membagi dataset menjadi 80% data latih dan 20% data uji.

e-ISBN 978-623-93343-4-5
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 15 September 2021

2.5 Modelling

2.5.1 Random Forest

Random forest merupakan algoritma dengan performance yang baik digunakan untuk melakukan klasifikasi data dalam jumlah yang besar. Dalam random forest akan melakukan suatu penggabungan data dengan decision tree dan juga untuk melakukan proses seleksi [7]. Sehingga saat penentuan klasifikasi pada decision tree, akan dilakukan pemecahan (split) berdasarkan jenis fitur yang ada dalam dataset. Dalam decision tree akan melakukan prediksi secara acak, sehingga dapat menghasilkan jawaban yang baik.

2.5.2 XGBoost

Extreme Gradient Boosting (XGBoost) adalah metode boosting dengan cara menggabungkan kumpulan pohon keputusan yang akan digunakan untuk pembangunan pohon selanjutnya [8]. Model ini dibangun dengan menggunakan metode boosting, yaitu dengan cara membuat model yang baru berdasarkan model yang lama, dan akan menghasilkan kesalahan prediksi (error) yang lebih kecil dari model sebelumnya. XGBoost adalah varian lain dari GBM yang lebih efisien dan terukur, XGBoost ini mampu menyelesaikan permasalahan seperti regresi, rangking, dan klasifikasi [9].

2.6 Evaluasi

Pada tahapan evaluasi, dilakukan pengujian model algoritma klasifikasi Random Forest, XGBoost dan Decision Tree hasil dari pembuatan model training dengan menggunakan data uji. Selanjutnyauntuk mengetahui efisiensi kinerja model yang telah dibuat, digunakan Confusion Matrix. Confusion Matrix merupakan pengukuran performa untuk suatu kasus klasifikasi atau metode pengukuran kinerja suatu model klasifikasi dengan 4 representasi yakni True Positif, True Negatif, False Positif, dan False Negatif [10]. Dengan confusion matrix dapat diketahui seberapa banyak data yang salah diklasifikasi pada data uji.

3. Hasil dan Pembahasan

Penelitian ini menggunakan data Weather AUS yang berasal dari situs Kaggle.com dengan format data csv. Dataset tersebut memiliki 23 fitur dengan jumlah data sebanyak 145.460. Penjelasan lebih lanjut mengenai atribut dalam dataset weather AUS dapat dilihat pada tabel 1.

Tabel. 1. Atribut variabel dataset.

Nama Atribut	Keterangan
Date	Tanggal Observasi
Location	Lokasi
MinTemp	Min suhu dalam celcius
MaxTemp	Max suhu dalam celcius
Rainfall	Jumlah Hujan
Evaporation	Evaporasi
Sunshine	Jumlah jam dalam hari
WindGustDir	Arah hembusan angin terkuat dalam 24 jam hingga tengah malam

WindGustSpeed	Kecepatan (km/jam) hembusan angin terkuat dalam 24 jam hingga tengah malam
WindDir9am	Arah angin pada jam 9 pagi
WindDir3pm	Arah angin pada jam 3 malam
WindSpeed9am	Kecepatan angin (km/jam) rata-rata lebih dari 10 menit sebelum jam 9 pagi
WindSpeed3pm	Kecepatan angin (km/jam) rata-rata lebih dari 10 menit sebelum jam 3 malam
Humidity9am	Kelembaban (persen) pada jam 9 pagi
Humidity3pm	Kelembaban (persen) pada jam 3 malam
Pressure9am	Tekanan atmosfer (hpa) berkurang menjadi rata-rata permukaan laut pada pukul 9 pagi
Pressure3pm	Tekanan atmosfer (hpa) berkurang menjadi rata-rata permukaan laut pada pukul 3malam
Cloud9am	Pecahan langit tertutup awan pada pukul 9 pagi
Cloud3pm	Bagian langit yang tertutup awan (dalam "oktas": seperdelapan) pada jam 3 sore
Temp9am	Suhu (derajat C) pada jam 9 pagi
Temp3pm	Suhu (derajat C) pada jam 3 malam
RainToday	Hujan Hari ini
RainTomorrow	Hujan Besok

3.1 Pre-processing data

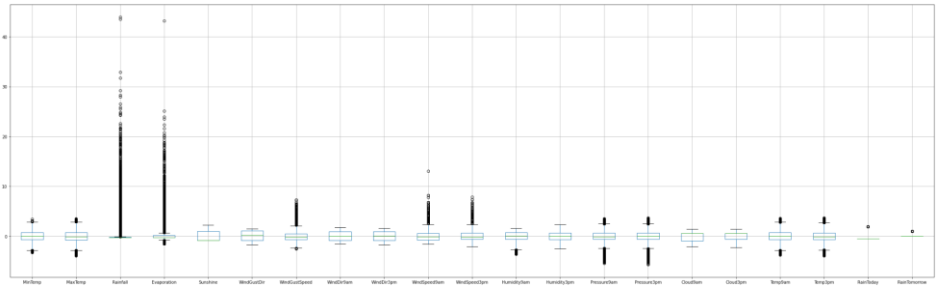
Setelah dilakukan eksplorasi data, ada beberapa tahapan yang harus dilakukan sebelum dilakukan klasifikasi model, tahapan yang dilakukan yaitu:

1. **Mengisi missing value.** pada visualisasi missing value pada proses eksplorasi data terlihat masih banyak nilai null/ missing value dikarenakan data weatherAUS tidak sepenuhnya bersih dan memiliki banyak noise, lalu dilakukan pengisian missing value dengan menggunakan data modus pada tiap variabelnya. Lalu dilakukan pengecekan missing value lagi untuk melihat apakah masih ada missing value.
2. **Mengubah tipe data.** Pada tahap ini variabel - variabel fitur yang masih memiliki tipe data objek seperti, WindGustDir, WindDir9am, WindDir3pm, RainToday dan RainTomorrow menjadi integer berdasarkan jenis data yang ada di dalam variabel tersebut.
3. **Menghapus fitur data.** Dalam dataset WeatherAUS, tidak semua variabel memiliki peran penting, seperti variabel date dan location yang tidak berpengaruh dalam proses klasifikasinya.
4. **Standarisasi.** Proses ini merubah isi dari data dalam variabel sehingga distribusinya akanmemiliki nilai rata-rata 0 dan standar deviasi 1. Data yang sudah dilakukan standarisasi dapat dilihat pada gambar 2.

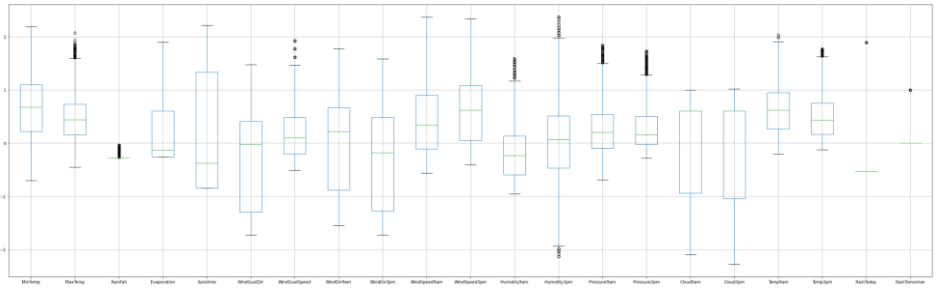
	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	I
0	0.191328	-0.041360	-0.203581	-0.257221	-0.845389	1.045228	
1	-0.751052	0.268745	-0.275097	-0.257221	-0.845389	1.258262	
2	0.112796	0.353318	-0.275097	-0.257221	-0.845389	1.471296	
3	-0.468338	0.677518	-0.275097	-0.257221	-0.845389	-0.872075	
4	0.835287	1.283631	-0.155903	-0.257221	-0.845389	1.045228	

Gambar. 2. Hasil Standarisasi Data

5. **Menghapus outlier.** Sebelum dilakukan penghapusan data pada outlier dilakukan visualisasi outlier yang terdapat pada gambar 3. Setelah itu dilakukan penghapusan data outlier terhadap data yang memiliki nilai persentil lebih dari 75 dan kurang dari 25. Lalu dilakukan visualisasi outlier lagi, untuk melihat hasil dari penghapusan outlier. Visualisasi dapat dilihat pada gambar 4.



Gambar. 3. Boxplot sebelum Penghapusan Outlier



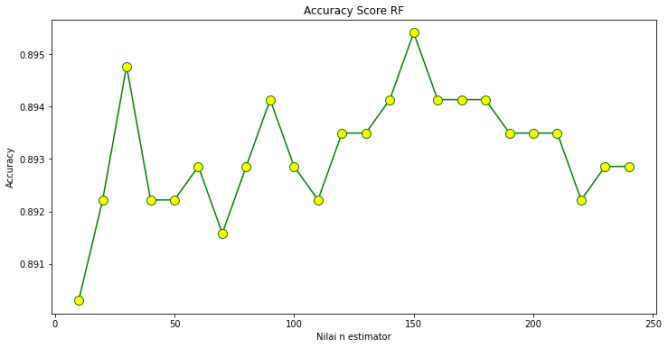
Gambar. 4. Boxplot setelah Penghapusan Outlier

3.2 Pembagian data

Setelah dilakukan tahapan praproses dihasilkan sebanyak 7.839 data yang kemudian dilakukan tahapan split data tanpa resampling, menghasilkan data latih sebanyak 6.271 data dan data uji sebanyak 1.568 data. Untuk tahapan split data dengan resampling, menghasilkan data latih sebanyak 11.168 data dan data uji sebanyak 2.792 data dari total 13.960 data.

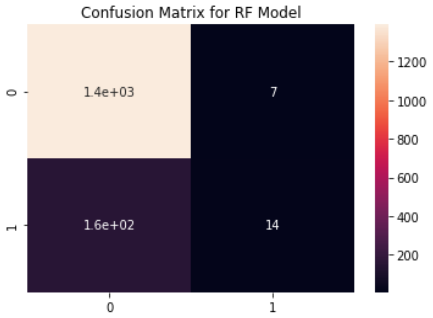
3.3 Random Forest

Pada proses klasifikasi random forest, dilakukan pengujian terhadap n_estimator dari range 10 hingga 250 dengan kelipatan 10. Hasil dari pengujian n_estimator dapat dilihat pada gambar 5.



Gambar. 5. Grafik Akurasi Random Forest

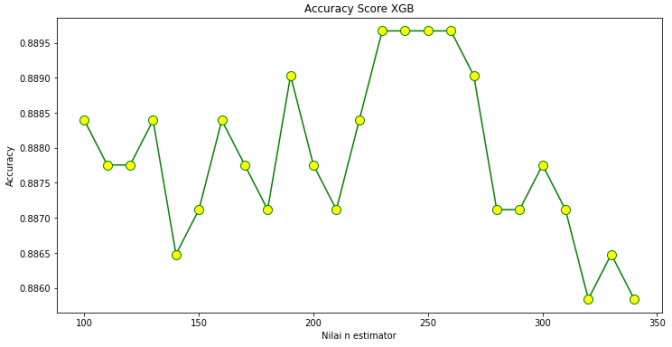
Setelah didapat n_estimator terbaik, yaitu 150 dengan nilai akurasi sebesar 89,54%. Lalu dilakukan evaluasi dengan menggunakan confusion matrix, dapat dilihat pada gambar 6. Hasil evaluasi klasifikasi random forest yaitu True Positive sebesar 1400, False Positive sebesar 7, False Negative sebesar 160 dan True Negative sebesar 14.



Gambar. 6. Hasil evaluasi klasifikasi Random Forest

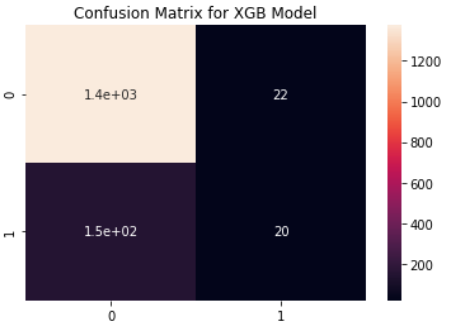
3.4 XGBoost

Pada proses klasifikasi XGBoost, dilakukan pengujian terhadap n_estimator dari range 10 hingga 350 dengan kelipatan 10. Hasil dari pengujian n_estimator dapat dilihat pada gambar 7.



Gambar. 7. Grafik Akurasi XGBoost

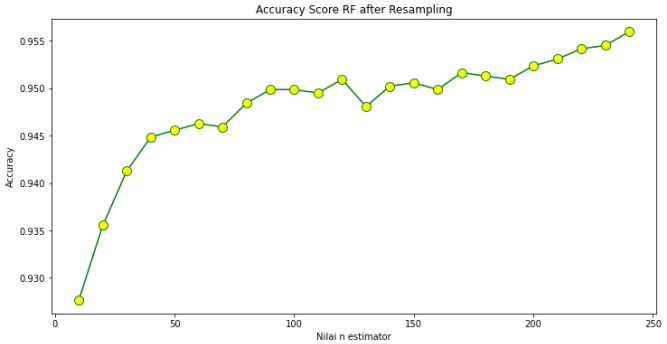
Setelah didapat n_estimator terbaik, yaitu 230 dengan nilai akurasi sebesar 88,96%. Lalu dilakukan evaluasi dengan menggunakan confusion matrix, dapat dilihat pada gambar 8. Hasil evaluasi klasifikasi XGBoost yaitu True Positive sebesar 1400, False Positive sebesar 22, False Negative sebesar 150 dan True Negative sebesar 20.



Gambar. 8.Hasil evaluasi klasifikasi XGBoost.

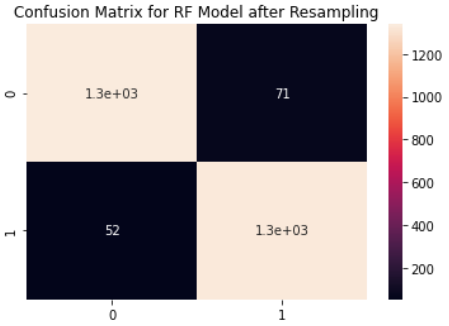
3.5 SMOTE Random Forest

Dengan menggunakan algoritma SMOTE, data klasifikasi No dan Yes pada variabel RainTomorrow menjadi sama yaitu 6980 data. Setelah itu dilakukan proses klasifikasi random forest terhadap n_estimator dari range 10 hingga 250 dengan kelipatan 10. Hasil dari pengujian n_estimator dapat dilihat pada gambar 9.



Gambar. 9. Grafik Akurasi SMOTE Random Forest

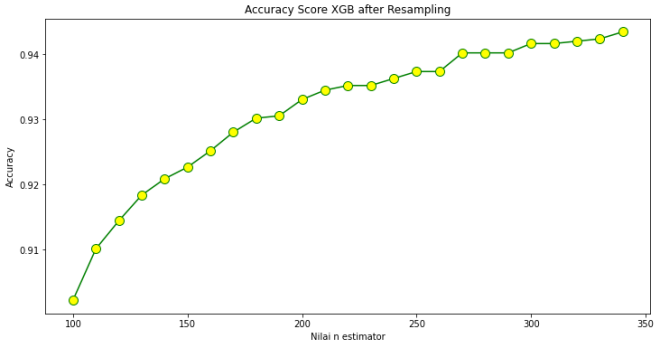
Setelah didapat n_estimator terbaik, yaitu 240 dengan nilai akurasi sebesar 95,59%. Lalu dilakukan evaluasi dengan menggunakan confusion matrix, dapat dilihat pada gambar 10. Hasil evaluasi klasifikasi SMOTE Random Forest yaitu True Positive sebesar 1300, False Positive sebesar 71, False Negative sebesar 52 dan True Negative sebesar 1300.



Gambar. 10. Hasil evaluasi klasifikasi SMOTE Random Forest

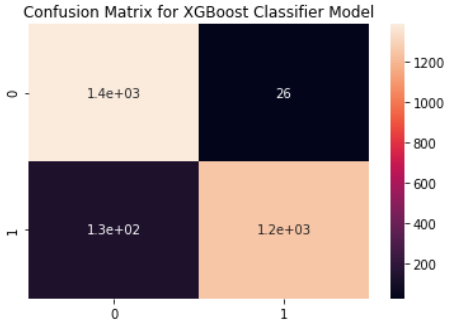
3.6 SMOTE XGBoost

Dengan menggunakan algoritma SMOTE, data klasifikasi No dan Yes pada variabel RainTomorrow menjadi sama yaitu 6980 data. Setelah itu dilakukan proses klasifikasi XGBoost terhadap n_estimator dari range 100 hingga 350 dengan kelipatan 10. Hasil dari pengujian n_estimator dapat dilihat pada gambar 11.



Gambar. 11. Grafik Akurasi SMOTE XGBoost

Setelah didapat n_estimator terbaik, yaitu 340 dengan nilai akurasi sebesar 94,34%. Lalu dilakukan evaluasi dengan menggunakan confusion matrix, dapat dilihat pada gambar 12. Hasil evaluasi klasifikasi SMOTE XGBoost yaitu True Positive sebesar 1400, False Positive sebesar 26, False Negative sebesar 130 dan True Negative sebesar 1200.



Gambar. 12. Hasil evaluasi klasifikasi SMOTE XGBoost

3.7 Perbandingan Metode Klasifikasi

Berdasarkan hasil evaluasi dari keempat metode klasifikasi yaitu random forest, XGBoost, SMOTE random forest, dan SMOTE XGBoost mendapatkan nilai akurasi lebih dari 80%, namun hasil recall dengan menggunakan SMOTE berbeda. Dengan menggunakan algoritma SMOTE membuat nilai recall random forest dan XGBoost meningkat menjadi lebih dari 90%

Tabel. 2. Perbandingan Akurasi dan Recall

	Random Forest	XGBoost	SMOTE RF	SMOTE XGB
Akurasi	89,54%	88,96%	95,59%	94,34%
Recall	8,18%	11,69%	96,23%	90,44%

4. Kesimpulan dan Saran

4.1 Kesimpulan

Hasil akhir pada penelitian ini menunjukkan bahwa akurasi tertinggi diperoleh klasifikasi Random Forest dengan Resampling yakni sebesar 95.59%, namun pada klasifikasi tanpa resampling akurasitertinggi diperoleh Random Forest yakni sebesar 89.54%. Pada penelitian ini membuktikan bahwa resampling menggunakan SMOTE dapat meningkatkan akurasi pada klasifikasi dan juga dapat meningkatkan recall pada proses klasifikasi. Jika dibandingkan dengan hasil akurasi klasifikasi tanpa resampling, akurasi dari random forest meningkat sebesar 6.05% sedangkan XGBoost meningkat sebesar 5.38%. Recall pada klasifikasi random forest meningkat sebesar 88.05% sedangkan XGBoost meningkat sebesar 78.75% hal ini menunjukkan bahwa dengan menggunakan resampling SMOTE, kelas dapat dikenali dengan baik.

4.2 Saran

Saran dari peneliti adalah perlunya tinjauan lebih lanjut mengenai metode klasifikasi lain, dikarenakan masih terdapat error sebesar 4.41%, metode klasifikasi lain berpotensi memiliki hasil akurasi yang lebih baik bahkan jika tidak menggunakan resampling. Diperlukan juga tinjauan lebih lanjut mengenai metode pre processing lain untuk meningkatkan akurasi.

5. Referensi

[1] Setiawan, P., Hidayat, A., & Sugiharto, T. (2010). ESTIMASI AIR MAMPU CURAH MENGGUNAKAN DATA MODIS SEBAGAI INFORMASI CUACA SPASIAL DI PULAU JAWA. *Jurnal Penginderaan Jauh dan Pengolahan Data Citra Digital*, 3(1).

[2] Dhawangkhara, M., & Riksakomara, E. (2017). Prediksi Intensitas Hujan Kota Surabaya dengan Matlab menggunakan Teknik Random Forest dan CART (Studi Kasus Kota Surabaya). *Jurnal Teknik ITS,* 6(1), 88–93.

[3] Siregar, A. M., dkk. (2020). PERBANDINGAN ALGORITME KLASIFIKASI UNTUK PREDIKSI CUACA. *Accounting Information System*, 15-24.

[4] Karo, I. M. K. (2020). Implementasi Metode XGBoost dan Feature Importance untuk Klasifikasi pada Kebakaran Hutan dan Lahan. *Journal of Software Engineering, Information and Communication Technology, 1*, 10-16.

[5] Sutoyo, E., & Fadlurrahman, M. A. (2020). Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network. *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, 6(3), 379-385.

[6] Affendi, F. M., dkk. (2013). Penerapan Synthetic Minority Oversampling Technique (Smote) Terhadap Data Tidak Seimbang Pada Pembuatan Model Komposisi Jamu. *Xplore, Vol. 1(1):e9(1-6)*.

[7] Renata, E., & Ayub, M. (2020). Penerapan Metode Random Forest untuk Analisis Risiko pada dataset Peer to peer lending. *Jurnal Teknik Informatika dan Sistem Informasi,* 462-474.

[8] Syukron, M., dkk. (2020). Perbandingan Metode Smote Random Forest Dan Smote Xgboost Untuk Klasifikasi Tingkat Penyakit Hepatitis C. *Jurnal Gaussian*, 227-236.

[9] Rachmi, A.N., (2020). Implementasi Metode Random Forest Dan Xgboost Pada Klasifikasi Customer Churn. *Laporan Tugas Akhir. Fakultas Matematika Dan Ilmu Pengetahuan Alam Universitas Islam Indonesia Yogyakarta.*

[10] Fauzi, A., dkk. (2020). Deteksi Penyakit Kanker Payudara dengan Seleksi Fitur berbasis Principal Component Analysis dan Random Forest. *Jurnal Infortech*, 96-101.