KLASIFIKASI CUACA PROVINSI DKI JAKARTA MENGGUNAKAN ALGORITMA RANDOM FOREST DENGAN TEKNIK OVERSAMPLING

Faqih Hamami[1)], Iqbal Ahmad Dahlan[2)]

[1]Fakultas Rekayasa Industri, Universitas Telkom
[2]Fakultas Teknik Militer, Universitas Pertahanan Indonesia
[1]Jl. Telekomunikasi No. 1, Terusan Buahbatu, Bandung
[2]Kawasan IPSC Sentul Sukahati, Bogor
Email: [1]faqihhamami@telkomuniversity.ac.id, [2]iqbal.dahlan@idu.ac.id

Abstract

Nowadays Indonesia experiences extreme weather changes that cause many disasters such as floods, fires, landslides and storms. The type of weather depends on many factors such as temperature, humidity, wind direction and others. Some human activities depend on changes in weather such as in the agricultural sector, plantations, aviation, highlands and beaches. Weather prediction is important to understand extreme weather changes based on weather factors. This research adopts ensemble learning which is able to perform weather classification. The algorithm used is Random Forest combined with oversampling technique to handle the uneven amount of data from each weather class. Some of the weather categories classified are Sunny, Sunny Cloudy, Cloudy, Heavy Cloudy, Local Rain, Light Rain, Moderate Rain and Thunderstorms. The experimental results show that the Random Forest model achieves accuracy of 70%. The oversampling technique used is the Synthetic Minority Over-sampling Technique (SMOTE) method. With the combination of SMOTE the predictions of each minority class can be increased by average of 50%..

Keyword: *classification, weather, Random Forest, oversampling, SMOTE*

Abstrak

Saat ini Indonesia sering mengalami perubahan cuaca ekstrem yang menyebabkan banyak bencana seperti banjir, kebakaran, longsor dan badai. Jenis cuaca bergantung dari banyak faktor seperti suhu, kelembaban, arah angin dan lainnya. Beberapa kegiatan manusia bergantung terhadap perubahan cuaca seperti di sektor pertanian, perkebunan, penerbangan, daerah tinggi dan pantai. Prediksi cuaca menjadi penting untuk lebih memahami perubahan cuaca ekstrem yang didasarkan dari faktor cuaca. Penelitian ini mengadopsi *ensemble learning* yang mampu melakukan klasifikasi cuaca dengan baik. Algoritma yang digunakan adalah *Random Forest* yang dikombinasikan dengan teknik *oversampling* untuk menangani ketidakmerataan jumlah data dari setiap kelas cuaca. Beberapa kategori cuaca yang diklasifikasikan adalah Cerah, Cerah Berawan, Berawan, Berawan Tebal, Hujan Lokal, Hujan Ringan, Hujan Sedang dan Hujan Petir. Hasil eksperimen diperoleh bahwa model *Random Forest* mencapai akurasi 70%. Teknik *oversampling* yang digunakan adalah metode *Synthetic Minority Over-sampling Technique* (SMOTE). Dengan kombinasi SMOTE prediksi dari setiap kelas minoritas dapat ditingkatkan dengan rata-rata sebesar 50%.

Kata Kunci: *klasifikasi, cuaca, Random Forest, oversampling, SMOTE*

1. Pendahuluan

Cuaca menjadi faktor penting dalam segala aspek kehidupan. Perubahan cuaca yang signifikan berpengaruh terhadap kegiatan manusia. Beberapa aktifitas yang berhubungan dengan cuaca seperti aktifitas pertanian, perkebunan dan penerbangan. Cuaca menjadi penting untuk dimonitor sehingga potensi seperti hujan deras, petir dapat dipersiapkan solusinya.

Perubahan cuaca bergantung pada banyak faktor seperti suhu, kelembaban, angin, waktu, lokasi dan lainnya. Dari faktor ini dapat diperoleh jenis cuaca seperti cuaca cerah, berawan, hujan, hujan petir. Cuaca ekstrim berpotensi menjadi bencana seperti banjir, tanah longsor, kebakaran, penyebaran penyakit yang mempengaruhi daya tahan tubuh manusia. Dengan teknik klasifikasi cuaca yang baik dapat diprediksi kemungkinan perubahan cuaca yang terjadi dengan lebih akurat.

Klasifikasi adalah pendekatan *supervised learning* yang beberapa algoritmanya seperti *Naïve Bayes, Decision Tree, Neural Network, SVM, Random Forest* [1][2]. Kasus klasifikasi dapat ditemukan diberbagai domain [3]–[5]. Salah satu pemanfaatan dalam klasifikasi adalah penentuan jenis cuaca. Beberapa penelitian mengimplementasikan algoritma pembelajaran mesin untuk klasifikasi dan prediksi cuaca [6][7]. [6] melakukan klasifikasi cuaca dengan beberapa algoritma klasifikasi seperti *Naïve Bayes, Decision Tree* dan *Random Forest*. [8] mengklasifikan cuaca dengan algoritma C4.5. Penelitian lain juga yang melakukan klasifikasi curah

hujan yang merupakan faktor penentu cuaca [9]. Perubahan cuaca dapat memicu bencana. [10] melakukan klasifikasi indeks cuaca kebakaran berdasarkan faktor cuaca menggunakan algoritma KNN.

Salah satu permasalahan dalam klasifikasi adalah persebaran data yang tidak merata. Jika data tidak tersebar merata di setiap kelas maka kemungkinan terjadinya *false negative* / *false positive* cukup tinggi di kelas minoritas. Kelas minoritas merupakan kelas dengan jumlah data yang lebih sedikit. Kondisi ini dapat mengakibabkan *classifier* belajar sedikit dari kelas minoritas sehingga prediksi akan lebih mengarah ke kelas mayoritas.

Penelitian ini mengusulkan pengembangan model *classifier* cuaca dengan pendekatan *ensemble learning* untuk pengkategorian cuaca berdasarkan atribut pendukung seperti lokasi, waktu, suhu dan kelembaban. Pengembangan model juga didukung untuk menangani ketimpangan jumlah data dari setiap kelas cuaca sehingga prediksi terhadap kelas minoritas menjadi lebih baik.

2. Metodologi

Random Forest merupakan salah satu *ensamble learning* yang dibangun dari pohon keputusan [2]. Beberapa keuntungan menggunakan pendekatan *ensamble learning* adalah dapat digunakan untuk kasus klasifikasi dan regresi, mampu memperoleh akurasi tinggi, cocok untuk analisis ukuran dataset yang besar dengan banyak dimensi. Selain itu *ensamble learning* seperti *Random Forest* juga cocok digunakan untuk menangani *imbalanced data* [11].

Ada beberapa pendekatan untuk menangani *imbalance data* yang salah satunya adalah dengan pendekatan *sampling* [12][13]. Metode *oversampling* merupakan bagian dari metode *sampling* yang popular untuk mengatasi *imbalanced data* [14]. *Oversampling* akan menambahkan jumlah data di kelas minoritas agar mempunyai jumlah yang sama dengan kelas mayoritas atau mendekatinya. Kombinasi algoritma *Random Forest* dengan metode *oversampling* diharapkan dapat menciptakan model pembelajaran yang lebih akurat dari *unbalanced data* [15][16].

Objek penelitian ini adalah kondisi cuaca di Provinsi DKI Jakarta yang datasetnya diambil melalui open data Jakarta. Periode waktu adalah di bulan Januari - Desember 2018. Dataset mempunyai 6 atribut yaitu *tanggal, wilayah, waktu, cuaca*, *kelebaban* dalam prosentase dan *suhu* dalam satuan *celcius* dengan total data sebanyak 8400. Detail atribut dari dataset dijelaskan pada Tabel 1.

Tabel 1. Detail Atribut Dataset

Nama Atribut	Keterangan
Tanggal	Tanggal data kelembaban dan suhu diambil
Wilayah	Wilayah data kelembaban dan suhu diambil
Waktu	Waktu data kelembaban dan suhu diambil
Cuaca	Jenis cuaca
Kelembaban	Nilai kelembaban dalam prosentase
Suhu	Nilai suhu dalam derajat celcius

Sedangkan potongan dari dataset cuaca di DKI Jakarta dapat dilihat pada Tabel 2.

Tabel 2. Preview Dataset Cuaca DKI Jakarta

Tanggal	2018-09-30	2018-09-30
Wilayah	Jakarta Barat	Jakarta Selatan
Waktu	Dini Hari	Dini Hari
Cuaca	Cerah Berawan	Berawan
Kelembaban	25-34^o	25-34^o
Suhu	50-58%	50-58%

Sebelum dataset siap digunakan untuk klasifikasi, diperlukan *preprocessing* data terlebih dahulu yang terdiri dari integrasi data, transformasi data dan pengkodean.

2.1. Data Integration

Data tersebar di 12 dataset berdasarkan bulan. Dataset perbulan harus diintegrasikan terlebih dahulu menjadi 1 dataset utuh untuk penyimpanan data dari bulan januari – desember. Metode untuk menggabungkan semua file adalah dengan menaruhnya di satu folder. Iterasi dilakukan untuk membaca setiap dataset dalam folder tersebut dan disimpan di dalam *list* untuk kemudian digabungkan.

2.2. Data Tranformation

Dari dataset diketahui kelas yang diprediksi adalah kolom Cuaca. Berdasarkan hasil eksplorasi, ditemukan bahwa terdapat 27 kelas yaitu *Cerah Berawan, Berawan, Hujan Lokal, Hujan Ringan, Cerah, Hujan Petir, Berawan Tebal, Hujan Sedang, Cerah Berawan, Cerah, Berawan, Hujan Lokal, Berawan, Cerah Berawah, Hujan, Hujan Ringan, Cerah berawan, Beawan, Hujan Ringanl, Cerang Berawan, Cerah, Hujan Loka, Hujang Sedang, Hujan Petir, Berawa, Hujan Sedang* dan *Cerah Berawn*.

Pada tabel cuaca diketahui terjadi perbedaan jenis cuaca tetapi merupakan entitas yang sama. Contohnya adalah "*Cerah Berawah*" adalah sama dengan "*Cerah Berawan*". Begitu juga "*Beawan*" dan "*Berawa*" merupakan jenis yang sama dengan "*Berawan*"

Diperlukan normalisasi kelas sehingga tidak terjadi duplikasi jenis cuaca. Dari 27 kelas cuaca akan dinormalisasi menjadi 8 kategori cuaca yaitu *Cerah, Cerah Berawan, Berawan, Berawan Tebal, Hujan Lokal, Hujan Ringan, Hujan Sedang* dan *Hujan Petir*. Detail preprocessing yang dilakukan pada kelas yang mempunyai makna sama ditampilkan di Table 3.

Tabel 3. Normalisasi Kelas Cuaca

Jenis Cuaca	Konversi Cuaca
Cerah berawan, Cerah Berawah, Cerang Berawan, Cerah Berawn	Cerah Berawan
Hujan, Hujan Ringan , Hujan Ringanl	Hujan Ringan
Berawan, Berawa, Beawan,	Berawan
Hujang Sedang	Hujan Sedang
Hujan Lokal , Hujan Loka	Hujan Lokal

Atribut *waktu* terdiri dari tahun, bulan dan tanggal. Atribut tanggal dipisah berdasarkan separator "-" dengan menggunakan ekspresi *lambda*. Perubahan bentuk *yyyy-mm-dd* menjadi "*yyyy*", '*mm*' dan '*dd*'.

Atribut *suhu* dan *kelembaban* mempunyai tipe data *string* dengan tambahan *postfix* % dan °. Nilai suhu dan kelembaban harus dikonversi menjadi bentuk numerik. Selain itu nilai suhu dan kelembaban berada dalam interval akan dipisah menjadi nilai minimal dan nilai maksimal.

Atribut waktu mempunyai 7 jenis waktu yaitu, *Dini Hari, Pagi, Siang, Malam, pagi, siang, malam*. Waktu "*Pagi*" dan "*pagi*" merupakan entitas yang merujuk kepada objek sama. Untuk mengatasinya dilakukan transformasi bentuk waktu menjadi *lowercase* sehingga dari 8 jenis waktu akan menjadi 4 yaitu *dini hari, pagi, siang, malam.*

2.3. Encoding

Sebelum data menjadi input dari algoritma pembelajaran mesin, setiap data katagorik yang berbentuk *string* harus dikonversi menjadi numerik. Ada 3 atribut yang bertipe data kategorik yaitu: *wilayah, waktu* dan *cuaca*.

One-hot encoding adalah adalah metode untuk mentransformasi data kategorik menjadi bentuk biner. Jika terdapat *n* kategori maka akan dikonversi menjadi *n* kolom dengan nilai biner. Nilai 1 merepresentasikan bahwa data tersebut memiliki nilai dengan kolom tadi.

One-hot encoding digunakan untuk menkonversi atribut wilayah dan waktu sedangkan atribut cuaca menggunakan metode *label encoding* yang merepresentasikan data kategorik menjadi bentuk angka desimal. Pendekatan ini dilakukan karena cuaca merupakan kelas target. *Encoding* dari jenis cuaca dapat dilihat pada Tabel 4.

Tabel 4. Pengkodean Kelas Cuaca

Kelas Cuaca	Jenis Cuaca
0	Berawan
1	Berawan Tebal
2	Cerah
3	Cerah Berawan
4	Hujan Lokal
5	Hujan Petir
6	Hujan Ringan
7	Hujan Sedang

2.4. Oversampling

Pada persoalan klasifikasi diperlukan pemisahan data menjadi 2 jenis, *training* dan *testing*. Data *training* digunakan untuk pembelajaran sedangkan data *testing* digunakan untuk menguji model yang dibangun. Pada dataset penelitian, jumlah data training pada setiap kelas tidak seimbang atau dikenal dengan istilah *imbalanced data*. Diperlukan pendekatan khusus sehingga klasifikasi yang dilakukan mempunyai akurasi yang tinggi untuk setiap kelas. Metode *oversampling* digunakan untuk mensintesis data di kelas yang mempunyai sedikit transaksi. Distribusi data di setiap kelas yang diurutkan berdasarkan jumlah datanya ditampilkan di Tabel 5.

Tabel 5. Distribusi Kelas Cuaca

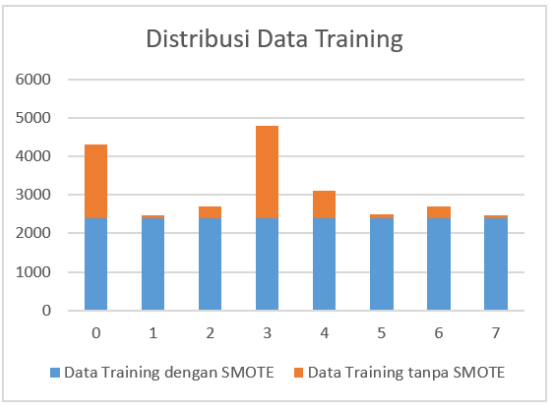
Kelas Cuaca	Jumlah Data
3	3345
0	2782
4	1020
6	452
2	436
5	138
7	115
1	112

Dari distribusi ini dapat dilihat bahwa kelas 1, 7 dan 5 hanya terdiri dari seratusan data. Untuk kelas 2 dan 6 mencapai empat ratusan data dan untuk kelas 0, 3 dan 4 mencapai ribuan data. Tentunya hal ini akan mengakibakan hasil klasifikasi kurang optimal dikarenakan distribusi data di setiap kelasnya berbeda jauh.

Diperlukan teknik *oversampling* untuk membangkitkan data baru di kelas minoritas seperti kelas 1,7,5 dan 2 sehingga proporsi jumlah data disetiap kelas sama atau berdekatan. Pada penelitian ini motede yang digunakan adalah Metode *Synthetic Minority Over-sampling Technique* (SMOTE). SMOTE mampu meningkatkan akurasi di model data dengan *imbalanced data*.

Jenis Data yang disentesis hanya data training. Data training merupakan dataset yang digunakan untuk pembelajaran. Ketika distribusi data tidak merata di data training maka akurasi dalam klasifikasi kelas minoritas tidak akan maksimal. Distribusi data

training yang dibangkitkan oleh SMOTE ditampilkan dalam visualisasi *stacked bar* di Gambar 1. Dari Gambar 1 dapat dipahami bahwa jumlah data training yang disintesis dari SMOTE mempunyai jumlah ukuran yang sama yaitu 2400 yang direpresentasikan dalam batang berwarna biru sedangkan batang berwarna orange menunjukkan distribusi jumlah asli dari setiap kelas cuaca.



Gambar 1. Perbandingan Jumlah Data *Training* sebelum dan sesudah diimplementasikan SMOTE

2.5. Modeling & Evaluation

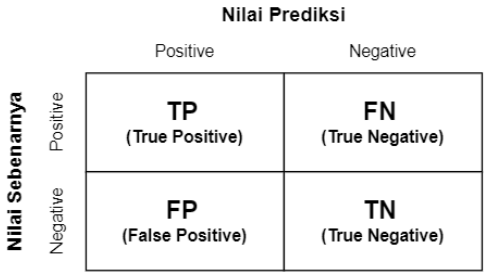
Algoritma klasifikasi yang digunakan adalah *Random Forest*. Algoritma ini menggunakan atribut *tanggal, wilayah, waktu, cuaca, kelembaban* dan *suhu* sebagai *independent variable* yang sudah dilakukan pembersihan serta transformasi data. Sedangkan atribut *cuaca* digunakan sebagai *dependen variable* atau kelas target yang menunjukkan jenis cuaca yang diklasifikasikan. Seperti kasus klasifikasi pada umumnya, dataset akan dipecah menjadi 2 yaitu sebagai data *training* dan *testing* dengan pembagian dataset di rasio 70:30. Data *training* disintenis dengan SMOTE sehingga jumlah data di setiap kelas adalah 2400 sedangkan data *testing* ditampilkan di Tabel 6.

Tabel 6. Pemisahan Data *Training* dan *Testing*

Kelas Cuaca	Jumlah Data	Data *Testing*
3	3345	945
0	2782	873
4	1020	311
6	452	128
2	436	154
5	138	38
7	115	32
1	112	39

Setelah data dipisah menjadi 2 bagian, data training digunakan untuk pembelajaran dengan menggunakan algoritma *Random Forest*. Pemodelan dibangun dengan 2 jenis yaitu pemodelan tanpa kombinasi SMOTE dan pemodelan dengan kombinasi dengan SMOTE.

Setelah modeling dilakukan, akurasi dari model akan diperoleh dengan menggunakan *confusion matrix*. *Confusion matrix* merupakan tabel khusus yang memvisualkan kinerja algoritma klasifikasi. Dari *confusion matrix* dapat diperoleh akurasi dari *true positive, true negative, false positive* dan *false negative.* Ilustrasi *confusion matrix* dapat dilihat di Gambar 2.



Gambar 2. *Confusion Matrix* untuk *Binary Classification*

Ilustrasi diatas adalah kasus *binary classification* atau klasifikasi terhadap 2 objek. Pada penelitian ini dibutuhkan *multi-class classification* karena kelas cuaca mempunyai 8 kelas. Karena bersifat *multi-class* maka jumlah kolom dan baris dari *confusion matrix* akan berjumlah *n* kelas cuaca.

Dari *confusion matrix* dapat diperoleh akurasi dari model. Akurasi menerangkan seberapa akurat model model learning dapat mengklasifikasikan cuaca. Akurasi dihitung dari rasio prediksi besar terhadap keseluruhan data. Akurasi model dapat diperoleh dari rumus ini.

$$\text{Accuracy} = \frac{TP+TN}{(TP+FP+TN+FN)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

dimana:
TF = *True Positive,* bernilai positif dan diprediksi positif
TN= *True Negative,* bernilai negatif dan diprediksi negatif
FP = *False Positive,* bernilai negatif dan diprediksi positif
FN= F*alse Negative,* bernilai positif dan diprediksi negatif

3. Hasil dan Pembahasan

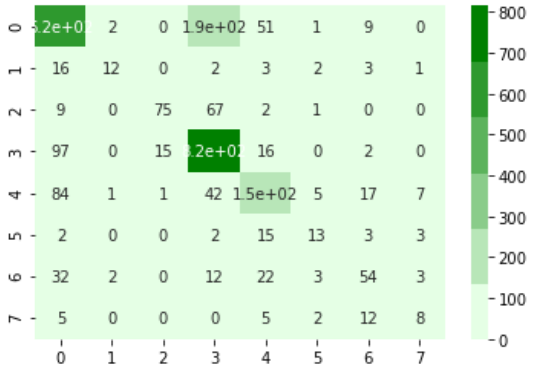
Ekperimen dilakukan dengan menguji data *testing* dari model *Random Forest* yang dilatih dengan kombinasi metode SMOTE dan tanpa SMOTE. Pemodelan data menggunakan *default hyperparameter* dari algoritma. Nilai dari *hyperparameter* yang digunakan di *Random Forest* adalah seperti di Tabel 7.

Tabel 7. Nilai *Hyperparameter Random Forest*

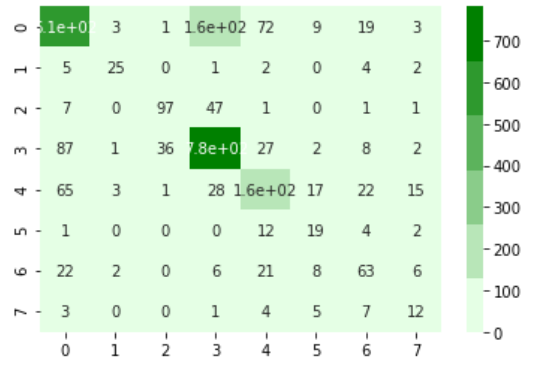
Hyperparameter	Nilai
n_estimators	100
criterion	‘gini’
max_depth	17
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_features	‘auto’
max_leaf_nodes	None
min_impurity_decrease	0.0

min_impurity_split	None
bootstrap	True
oob_score	False
n_jobs	None
random_state	0
verbose	0
warm_start	False
class_weight	None
ccp_alpha	0.0
max_samples	None

Berdasarkan ekperimen diperoleh akurasi algoritma *Random Forest* menggunakan SMOTE adalah 70% sedangkan tanpa SMOTE adalah 69%. Selisih antara kedua model ini sangat kecil. Walaupun akurasi hampir sama tetapi jika dilihat menggunakan *confusion matrix*, prediksi dari kelas minoritas dengan SMOTE dianggap lebih baik. Visualisasi *confusion matrix* dari model *Random Forest* menggunakan SMOTE dapat dilihat pada Gambar 3 sedangkan tanpa SMOTE dapat dilihat di Gambar 4.



Gambar 3. *Confusion matrix* dari *Random Forest* tanpa SMOTE



Gambar 4. *Confusion matrix* dari *Random Forest* dengan SMOTE

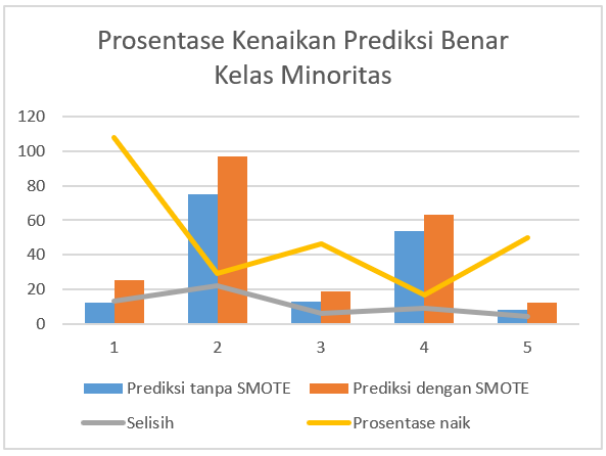
Dari kedua hasil modeling diperoleh *true positive*, *true negative*, *false positive* dan *false negative* dari setiap kelas. Fokus pada penelitian ini adalah pada kelas minoritas yaitu kelas 1, 2, 5, 6 dan 7. Pewarnaan dalam matriks adalah menunjukkan ukuran jumlah. Warna hijau muda menunjukkan nilai rendah dan semakin hijau tua maka jumlah prediksi cuaca semakin besar. Karena terdapat 8 kelas maka matriks membentuk 8 x 8 yang merepresentasikan jenis cuaca dalam kelas 0-7.

Pada kelas 1 jumlah data *testing* adalah 39 dengan prediksi benar model tanpa SMOTE adalah 12 sedangkan model dengan SMOTE adalah 25. Pada kelas 2 jumlah data *testing* adalah 154 dengan prediksi benar model tanpa SMOTE adalah 75 sedangkan model dengan SMOTE adalah 97. Pada kelas 5 jumlah data *testing* adalah 38 dengan prediksi benar model tanpa SMOTE adalah 13 sedangkan model dengan SMOTE adalah 19. Pada kelas 6 jumlah data *testing* adalah 128 dengan prediksi benar model tanpa SMOTE adalah 54 sedangkan model dengan SMOTE adalah 63. Pada kelas 7 jumlah data *testing* adalah 32 dengan prediksi benar model tanpa SMOTE adalah 8 sedangkan model dengan SMOTE adalah 12. Detail dari hasil *confusion matrix* untuk kelas 1, 2, 5, 6 dan 7 dapat dilihat pada Tabel 8.

Tabel 8. Hasil *Confusion Matrix* Kelas Minoritas

Kelas Cuaca	Data *Testing*	Prediksi tanpa SMOTE	Prediksi dengan SMOTE
1	39	12	25
2	154	75	97
5	38	13	19
6	128	54	63
7	32	8	12

Berdasarkan Table 8, prosentase kenaikan prediksi benar di kelas 1 naik sebesar 108%, Kelas 2 bertambah 29%, kelas 5 bertambah 46%, kelas 6 bertambah 16% dan kelas 7 bertambah 50%. Rata-rata kenaikan prediksi benar dari kombinasi SMOTE mencapai 50%. Detail kenaikan dari setiap kelas minoritas divisualkan dalam grafik batang di Gambar 5.



Gambar 5. Perbandingan Prediksi Benar, Selisih dan Kenaikan Prosentase di Kelas Minoritas

Grafik batang menunjukkan perbandingan antara prediksi benar dalam klasifikasi jenis cuaca. Batang biru adalah prediksi tanpa SMOTE dan batang orange adalah prediksi benar dengan SMOTE di kelas minoritas. Sumbu *y* pada grafik batang dalam satuan jumlah prediksi.

Grafik garis terdiri dari 2 bagian yaitu selisih prediksi benar dan prosentase kenaikannya. Garis berwarna abu menunjukkan selisih antara prediksi benar di setiap kelas sedangkan garis berwarna kuning merupakan prosentase kenaikan dari setiap kelas. Sumbu *y* pada grafik garis kuning dalam satuan prosentase. Kelas 1 mempunyai kenaikan paling tinggi dari kelas yang lain yaitu mencapai 108%.

4. Kesimpulan

Cuaca merupakan keadaan udara yang berubah-ubah bentuknya didasarkan dari banyak faktor. Klasifikasi cuaca penting bagi masyarakat yang bergerak dibidang pertanian, perkebunan, penerbangan serta diharapkan lebih memahani dampak lingkungan. Penelitian ini menggunakan algoritma *Random Forest* untuk mengklasifikasikan cuaca. Pembelajaran juga digabungkan dengan teknik *oversampling* untuk menangani jumlah data di kelas minoritas. Berdasarkan eksperimen diperoleh akurasi model *Random Forest* mencapai 70% dengan teknik SMOTE. Hasil klasifikasi juga mampu memperbaiki prediksi setiap kelas minoritas 1, 2, 5 , 6 dan 7 dengan rata-rata kenaikan prediksi benar mencapai 50%.

Daftar Pustaka

[1] S. Umadevi and K. S. J. Marseline, "A survey on data mining classification algorithms," *Proc. IEEE Int. Conf. Signal Process. Commun. ICSPC 2017*, vol. 2018-January, no. July, pp. 264–268, 2018, doi: 10.1109/CSPC.2017.8305851.

[2] A. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha, and S. Kundu, "Improved Random Forest for Classification," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4012–4024, 2018, doi: 10.1109/TIP.2018.2834830.

[3] R. Katari and M. B. Myneni, "A Survey on News Classification Techniques," *2020 Int. Conf. Comput. Sci. Eng. Appl. ICCSEA 2020*, pp. 8–12, 2020, doi: 10.1109/ICCSEA49143.2020.9132866.

[4] Y. Sun, Y. Li, Q. Zeng, and Y. Bian, "Application research of text classification based on random forest algorithm," *Proc. - 2020 3rd Int. Conf. Adv. Electron. Mater. Comput. Softw. Eng. AEMCSE 2020*, pp. 370–374, 2020, doi: 10.1109/AEMCSE50948.2020.00086.

[5] F. Hamami and I. Fithriyah, "Classification of air pollution levels using artificial neural network," *2020 Int. Conf. Inf. Technol. Syst. Innov. ICITSI 2020 - Proc.*, pp. 217–220, 2020, doi: 10.1109/ICITSI50517.2020.9264910.

[6] A. M. Siregar, "Klasifikasi Untuk Prediksi Cuaca Menggunakan Esemble Learning," *Petir*, vol. 13, no. 2, pp. 138–147, 2020, doi: 10.33322/petir.v13i2.998.

[7] N. Singh, S. Chaturvedi, and S. Akhter, "Weather Forecasting Using Machine Learning Algorithm," *2019 Int. Conf. Signal Process. Commun. ICSC 2019*, pp. 171–174, 2019, doi: 10.1109/ICSC45622.2019.8938211.

[8] A. Novandya, "Penerapan Algoritma Klasifikasi Data Mining C4.5 pada Dataset Cuaca Wilayah Bekasi," *KNiST*, vol. 6, pp. 368–372, 2017.

[9] I. G. A. Gunadi, A. Aprilyana, and K. Dewi, "Klasifikasi Curah Hujan di Provinsi Bali Berdasarkan Metode Naive Bayesian," *J. Mat. Sains, dan Pembelajarannya*, vol. 12, no. 1, pp. 14–25, 2018.

[10] M. Reza Noviansyah, T. Rismawan, and D. Marisa Midyanti, "Penerapan Data Mining Menggunakan Metode K-Nearest Neighbor Untuk Klasifikasi Indeks Cuaca Kebakaran Berdasarkan Data Aws (Automatic Weather Station) (Studi Kasus: Kabupaten Kubu Raya)," *J. Coding, Sist. Komput. Untan*, vol. 06, no. 2, pp. 48–56, 2018, [Online]. Available: http://jurnal.untan.ac.id/index.php/jcskommipa/article/view/26672.

[11] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 4, pp. 1–18, 2018, doi: 10.1002/widm.1249.

[12] V. S. Spelmen and R. Porkodi, "A Review on Handling Imbalanced Data," *Proc. 2018 Int. Conf. Curr. Trends Towar. Converging Technol. ICCTCT 2018*, pp. 1–11, 2018, doi: 10.1109/ICCTCT.2018.8551020.

[13] S. Yadav and G. P. Bhole, "Handling Imbalanced Dataset Classification in Machine Learning," *2020 IEEE Pune Sect. Int. Conf. PuneCon 2020*, pp. 38–43, 2020, doi: 10.1109/PuneCon50868.2020.9362471.

[14] A. S. More and D. P. Rana, "Review of random forest classification techniques to resolve data imbalance," *Proc. - 1st Int. Conf. Intell. Syst. Inf. Manag. ICISIM 2017*, vol. 2017-Janua, pp. 72–78, 2017, doi: 10.1109/ICISIM.2017.8122151.

[15] Y. Pristyanto, A. F. Nugraha, I. Pratama, A. Dahlan, and L. A. Wirasakti, "Dual Approach to Handling Imbalanced Class in Datasets Using Oversampling and Ensemble Learning Techniques," *Proc. 2021 15th Int. Conf. Ubiquitous Inf. Manag. Commun. IMCOM 2021*, 2021, doi: 10.1109/IMCOM51814.2021.9377420.

[16] H. Luo, X. Pan, Q. Wang, S. Ye, and Y. Qian, "Logistic regression and random forest for effective imbalanced classification," *Proc. - Int. Comput. Softw. Appl. Conf.*, vol. 1, pp. 916–917, 2019, doi: 10.1109/COMPSAC.2019.00139.