DOI: doi.org/10.21009/03.1201.FA18

ANALISIS MODEL PREDIKSI CUACA MENGGUNAKAN *SUPPORT VECTOR MACHINE, GRADIENT BOOSTING, RANDOM FOREST*, DAN *DECISION TREE*

Risanti[a)], Widyaningrum Indrasari[b)] , Haris Suhendar[c)]

Program Studi Fisika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta. Jl. Rawamangun Muka No 1, Jakarta Timur 13220, Indonesia.

Email: [a)]ri.santiii21@gmail.com, [b)]widyaningrum-indrasari@unj.ac.id, [c)]haris_suhendar@unj.ac.id

Abstrak

Machine learning dapat diaplikasikan untuk melakukan prediksi terhadap suatu data. Salah satu data yang berkaitan dengan fenomena alam yang terdokumentasi dengan baik dan dapat diakses dengan mudah adalah data kondisi cuaca. Dalam penelitian ini digunakan data kondisi cuaca untuk melakukan pengembangan model machine learning dan prediksi keadaan cuaca. Data yang digunakan terdiri dari pengukuran suhu udara, kelembaban udara, dan kecepatan angin menggunakan data BMKG Provinsi Jawa yang bersifat *open source* dengan selang waktu 3 jam tahun 2020 dari bulan Januari - Desember. Tujuan penelitian ini adalah untuk mendapatkan nilai akurasi, presisi, F1 *score*, dan *recall* serta membandingkan fitur yang memberikan pengaruh paling besar terhadap hasil prediksi cuaca. Model yang digunakan dalam penelitian ini adalah *support vector machine*, *gradient boosting*, *random forest*, dan *decision tree*. Perbandingan antara data training dan data test adalah 70:30. Hasil penelitian menunjukan bahwa hasil akurasi model *support vector machine*, *gradient boosting*, *random forest*, *decision tree* masing-masing sebesar 0.1697; 0.6696; 0.7918; 0.8416; 0.8280. Pada hasil terlihat bahwa *random forest* memiliki pengaruh paling besar terhadap hasil prediksi cuaca dengan dengan hasil akurasi 0.8416.

Kata-kata kunci: Prediksi, Cuaca, SVM, *Gradient Boosting*, *Random Forest*, *Decision Tree*.

Abstract

Machine learning can be applied to make predictions on a given dataset. One well-documented and easily accessible dataset related to natural phenomena is weather condition data. In this study, weather condition data is used to develop machine learning models and predict weather conditions. The data used consists of air temperature, air humidity, and wind speed measurements obtained from the BMKG (Meteorology, Climatology, and Geophysics Agency) of the Jawa Province, which are open source and collected at 3-hour intervals throughout the year 2020 from January to December. The aim of this research is to obtain accuracy, precision, F1 score, and recall values and compare the features that have the most significant influence on weather prediction outcomes. The models used in this study are support vector machine, gradient boosting, random forest, and decision tree. The data is divided into a 70% training set and a 30% test set. The research results show that the accuracy values for, support vector machine, gradient boosting, random forest, and decision tree models are 0.1697, 0.6696, 0.7918, 0.8416, and 0.8280, respectively. It can be observed that random forest has the greatest influence on weather prediction outcomes, with an accuracy value of 0.8416.

Keywords: Prediction, Weather, SVM, Gradient Boosting, Random Forest, Decision Tree.

PENDAHULUAN

Prediksi cuaca merupakan hal yang sangat penting berbagai bidang kehidupan manusia [1]. Kebutuhan prakiraan cuaca yang akurat akan efektif dan efisien dalam mengelola kualitas peradaban secara fleksibel. Untuk mendapatkan prediksi cuaca yang akurat dibutuhkan algoritma yang tepat dan akurat untuk menentukannya[2]. Algoritma *machine learning* secara luas digunakan untuk prediksi, dalam beberapa tahuhn terakhir. Menurut expert system pada tahun 2017, *machine learning* adalah bagian dari kecerdasan buatan (*artificial intelligence*) yang memiliki kemampuan untuk belajar secara otomatis dan dapat meningkatkan kemampuannya berdasarkan pengalaman tanpa diprogram secara eksplisit. *Machine learning* juga dapat didefinisikan sebagai algoritma yang bertujuan menemukan pola-pola dalam data [3]. Kemajuan dalam bidang *machine learning* memungkinkan untuk mendapatkan hasil prediksi secara efisien dan akurat. Model *machine learning* yang digunakan pada penelitian kali ini adalah *Support Vector Machines, Gradient boosting, Decision tree, Random forest*. Prakiraan cuaca sangat penting dalam berbagai bidang kehidupan manusia, termasuk di kota-kota besar.

Support Vector Machines (SVM) adalah salah satu model pengklasifikasian pada *machine learning* dengan model biner atau diskriminatif, bekerja pada dua kelas diferensiasi. SVM digunakan untuk mencari *hyperplane* terbaik dengan memaksimalkan jarak antar kelas[4]. *Gradient boosting* adalah salah satu algoritma klasifikasi pembelajaran mesin yang menggunakan *ensemble* pohon keputusan yang memprediksi suatu nilai. *Gradient boosting* merupakan teknik menghasilkan model prediksi aditif dengan menggabungkan berbagai prediktor lemah, seperti dec*ision tree. Gradient boosting* dapat digunakan untuk regresi dan klasifikasi[5]. *Decision tree* digunakan dalam klasifikasi dan prediksi. *Decision tree* termasuk ke dalam *supervised learning*. *Decision tree* digunakan untuk klasifikasi data dari himpunan data dalam membuat keputusan [6]. *Random forest* merupakan kumpulan pohon-pohon dari *decision tree*. *Random forest* adalah metode ensemble yang biasa digunakan dalam merata-ratakan hasil dari pohon keputusan (*decision tree*) untuk *klasifikasi*, regresi, dan bentuk pembelajaran mesin lainnya. Untuk klasifikasi, prediksi akhir dari *ensemble* diberikan dengan suara terbanyak [7].

Penelitian mengenai prakiraan cuaca telah dilakukan oleh M. Pratapa Raju dan A. Jaya Laxmi (2020) mengenai peramalan *online load* berbasis IoT menggunakan algoritma *machine learning*[8]. Hemalatha, dkk. (2021) mengembangkan prediksi cuaca menggunakan Teknik *machine learning* untuk membuktikan bahwa parameter input cuaca dan output kondisi cuaca memiliki hubungan tidak linier[9]. Fowdur, dkk. (2022) melakukan penelitian mengenai *A real-time collaborative machine learning based weather forecasting system with multiple predictor locations*[10].

Oleh karena itu, pada penelitian ini akan dilakukan analisis prediksi cuaca menggunakan model *machine learning*. Pada penelitian kali ini akan digunakan parameter suhu udara, kelembaban udara, dan kecepatan angin. Dalam pengumpulan data, penelitian ini menggunakan metode eksperimen pengujian alat prediksi curah hujan menggunakan model *machine learning*. Data yang digunakan terdiri dari pengukuran suhu udara, kelembaban udara, dan kecepatan angin menggunakan data BMKG Provinsi Jawa yang bersifat *open source* pada bulan Januari - Desember tahun 2020 dengan selang waktu 3 jam. Data akan di-training dan di-testing menggunakan 5 model yang sudah ditentukan hingga didapatkan hasil nilai akurasi, presisi, F1-*score, recall*, serta membandingkan fitur yang memberikan pengaruh paling besar terhadap hasil prediksi cuaca menggunakan grafik *feature importance*.

METODOLOGI

Penelitian ini menggunakan empat model *machine learning* yaitu, *support vector machine, gradient boosting, random forest, dan decision tree*. Penelitian ini menggunakan data cuaca sekunder dari BMKG (Badan Meteorologi Klimatologi dan Geofisika) yang bersifat *open souce.* Data cuaca ini diambil di Provinsi Jawa pada tahun 2020 dengan selang waktu 3 jam. Data yang digunakan dalam penelitian ini terdiri dari suhu udara minimum, suhu udara maksimum, kelembaban udara minimum, kelembaban udara maksimum, kecepatan angin dan kategorikal cuaca. Sebelum menguji data, perlu dilakukan *import library* dan *preparing* data yang dibutuhkan, pada penelitian ini

menggunakan *library sklearn*. Setelah itu, dilakukan *pre-processing* data seperti pengecekan *missing value*, dan penghapusan *outliers.* Pada *pre-processing* data, atribut cuaca dibagi menjadi 3 label yaitu label 0 dengan kategori cerah, label 1 dengan kategori hujan, label 2 dengan kateogori berawan. Selanjutnya, data yang telah dibersihkan di *split* menjadi data *train* dan data *test* dengan perbandingan 80:20. Setelah itu, data di uji menggunakan empat model yg sudah dipilih dengan *output* nya adalah akurasi, presisi, recall, *f1-score*. Hasil data yang telah diuji akan dibandingkan dengan grafik *feature importance*, untuk mengetahui fitur yang memberikan pengaruh paling besar terhadap hasil prediksi cuaca. Berikut adalah diagram alir metodologi.

GAMBAR 1. Diagram Alir Penelitian

HASIL DAN PEMBAHASAN

Pengujian model menggunakan data cuaca sekunder dari BMKG (Badan Meteorologi Klimatologi dan Geofisika) yang bersifat *open souce.* Data cuaca ini diambil di Provinsi Jawa pada tahun 2020 dengan selang waktu 3 jam. Berikut adalah informasi dari dataset nya.

TABEL 1. Load Dataset

	ID lokasi	Waktu	suhu min	suhu max	RH min	RH max	RH	suhu	Kode cuaca	arah angin	kec angin
0	5009714	1/5/2020 0:00	NaN	NaN	NaN	NaN	75	26	60	NW	10
1	5009714	1/5/2020 3:00	NaN	NaN	NaN	NaN	70	28	3	W	20
2	5009714	1/5/2020 6:00	NaN	NaN	NaN	NaN	60	33	3	W	30
3	5009714	1/5/2020 9:00	NaN	NaN	NaN	NaN	75	28	3	W	10
4	5009714	1/5/2020 12:00	23.0	33.0	60.0	95.0	85	26	60	W	10
...	...	...	...	...	...	...	...	...	...	...	...
8075	5009904	1/10/2020 18:00	NaN	NaN	NaN	NaN	90	23	1	W	10
8076	5009904	1/11/2020 0:00	NaN	NaN	NaN	NaN	95	24	3	N	0
8077	5009904	1/11/2020 6:00	NaN	NaN	NaN	NaN	85	27	80	NW	10
8078	5009904	1/11/2020 12:00	23.0	27.0	80.0	100.0	100	24	5	E	10
8079	5009904	1/11/2020 18:00	NaN	NaN	NaN	NaN	80	23	1	W	10

8080 rows × 11 columns

Pada hasil plot, data berukuran 8080 *rows* x 11 *columns*. Pada dataset ada beberapa data yang tidak digunakan, maka akan dilakukan *feature selection* serta menghilangkan data NaN. Fitur yang akan digunakan adalah Pada dataset BMKG di atas, terdapat beberapa atribut yang tidak dibutuhkan. Maka dari itu, dilakukan *feature selection* untuk mengambil atribut yang diperlukan seperti, suhu minumum, suhu maksimum, RH (*Relative Humidity*) minimum, RH (*Relative Humidity*) maksimum, RH (*Relative Humidity*), suhu, kode cuaca dan kecepatan angin. Berikut adalah hasilnya:

TABEL 2. Hasil plot Feature Selection

	suhu min	suhu max	RH min	RH max	kec angin	Kode cuaca
4	23.0	33.0	60.0	95.0	10	3.0
12	22.0	32.0	60.0	95.0	10	3.0
20	22.0	31.0	60.0	95.0	10	3.0
26	25.0	28.0	85.0	95.0	10	80.0
30	24.0	26.0	95.0	95.0	10	1.0
...	...	...	...	...	...	...
8060	24.0	31.0	65.0	95.0	20	3.0
8066	23.0	26.0	95.0	100.0	10	80.0
8070	23.0	25.0	85.0	100.0	10	80.0
8074	23.0	26.0	85.0	95.0	0	1.0
8078	23.0	27.0	80.0	100.0	10	1.0

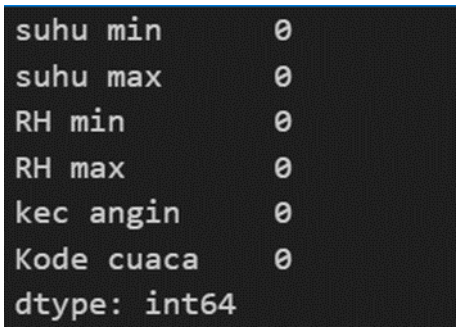
1414 rows × 6 columns

Pada informasi data df.info() terlihat bahwa tipe data suhu min, suhu max, RH min, RH max berbentuk float64 dan data kecepatan angin berbentuk int64 dengan memori yang digunakan sebesar 77.3 KB. Informasi dataset setelah dilakukan *pre-processing* data terdapat pada GAMBAR 2 berikut.

```
<class 'pandas.core.frame.DataFrame'>
Index: 1414 entries, 4 to 8078
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   suhu min    1414 non-null   float64
 1   suhu max    1414 non-null   float64
 2   RH min      1414 non-null   float64
 3   RH max      1414 non-null   float64
 4   kec angin   1414 non-null   int64  
 5   Kode cuaca  1414 non-null   float64
dtypes: float64(5), int64(1)
memory usage: 77.3 KB
```

GAMBAR 2. Hasil Plot Informasi Data

Setelah dilakukan *feature selection* dan menghilangkan data NaN, ukuran data berkurang menjadi 1414 *rows* × 6 *columns* dengan tipe data adalah float64 dan kecepatan angin berbentuk int64 dengan memori yang digunakan sebesar 77.3 KB. Selanjutnya dilakukan pengecekkan *missing value* pada setiap kolom. Berikut hasil plot pada GAMBAR 3.



GAMBAR 3. Pengecekan *Missing Value*

Pada hasil plot tidak terdeteksi adanya *missing value.* Kategori label dilakukan untuk mengurangi label pada atribut kode cuaca yang semula berjumlah 14 label menjadi 3 label. Label BMKG terdiri dari label 0 untuk cerah, label 1 untuk cerah berawan, label 2 untuk cerah berawan, label 3 untuk berawan, label 4 untuk berawan tebal, label 5 untuk udara kabur, label 10 untuk asap, label 45 untuk kabut, label 60 untuk hujan ringan, label 61 untuk hujan sedang, label 63 untuk hujan lebat, label 80 untuk hujan lokal, label 95 untuk hujan petir, label 97 untuk hujan petir dan label dipangkas menjadi label 0 untuk cerah, label 1 untuk hujan, dan label 2 untuk berawan. Pada dataset juga dihilangkan outliers nya. Dataset menjadi berukuran 1207 rows x 6 columns yang tertera pada TABEL 3 berikut:

TABEL 3. Hasil plot Pengkategorian Label dan Outliers

	suhu min	suhu max	RH min	RH max	kec angin	Cuaca
4	23.0	33.0	60.0	95.0	10.0	0
12	22.0	32.0	60.0	95.0	10.0	0
20	22.0	31.0	60.0	95.0	10.0	0
26	25.0	28.0	85.0	95.0	10.0	2
30	24.0	26.0	95.0	95.0	10.0	0
...	...	...	...	...	...	...
8052	24.0	31.0	65.0	95.0	20.0	0
8060	24.0	31.0	65.0	95.0	20.0	0
8066	23.0	26.0	95.0	100.0	10.0	2
8070	23.0	25.0	85.0	100.0	10.0	2
8078	23.0	27.0	80.0	100.0	10.0	0

1207 rows × 6 columns

Setelah dilakukan *pre-processing* dan *cleaning*, ukuran data menjadi 1207 *rows* x 6 *columns.* Selanjutnya, dilakukan evaluasi metrik. Evaluasi berfungsi untuk mengetahui akurasi dari model algoritma yang dibuat[11]. Kriteria evaluasi yang dipertimbangkan adalah akurasi, standard deviation, f1 score, recall, precision dan specificity [12] [13]. Pada penelitian ini, evaluasi yang dilakukan adalah dengan menghitung akurasi presisi, recall dan f1 score. Pada pengujian ini data menjadi data latih dan data uji dengan perbandingan 80:20. Pengujian dilakukan untuk mendapatkan hasil akurasi, presisi, recall, dan F-1 score serta hasilnya dibandingkan menggunakan grafik *feature importance* untuk mengetahui fitur mana yang paling berpengaruh terhadap prediksi cuaca.

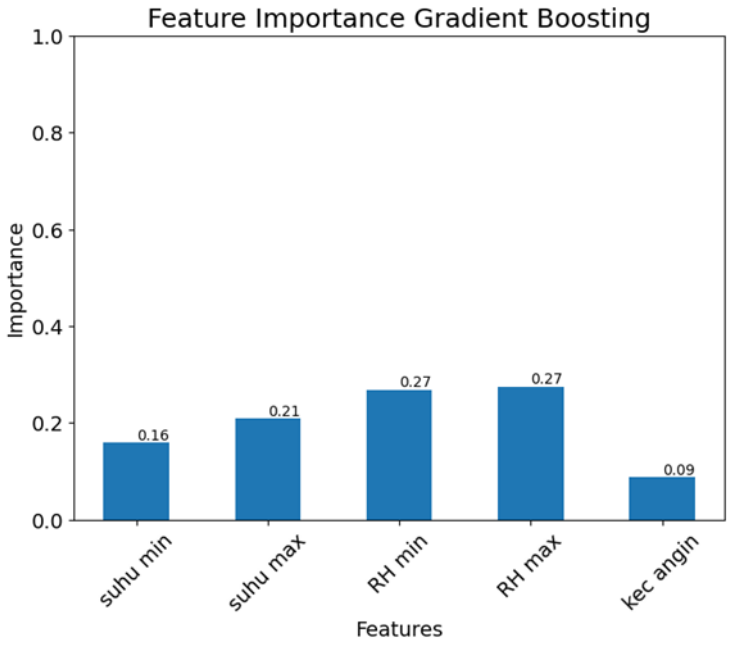
```
Model SVM
Accuracy: 0.6776859504132231
Precision: 0.6606060606060606
Recall: 0.5466045066045065
F1-score: 0.5772067176322495
```

GAMBAR 4. Hasil Metrik Evaluasi Umum pada Model *Support Vector Mahchine*

Pada hasil evaluasi metrik terlihat bahwa model *support vector machine* memiliki akurasi yang cukup rendah yaitu sebesar $\pm$ 67% dengan presisi ± 66%, dalam hal ini dapat dikatakan bahwa model tidak cocok dengan data karena memiliki akurasi yang terbilang rendah. Sedangkan pada model *gradient boosting* memiliki hasil sebagai berikut.

```
Model gradient boosting
Accuracy: 0.8181818181818182
Precision: 0.8164160401002506
Recall: 0.770034965034965
F1-score: 0.7905194552184617
```

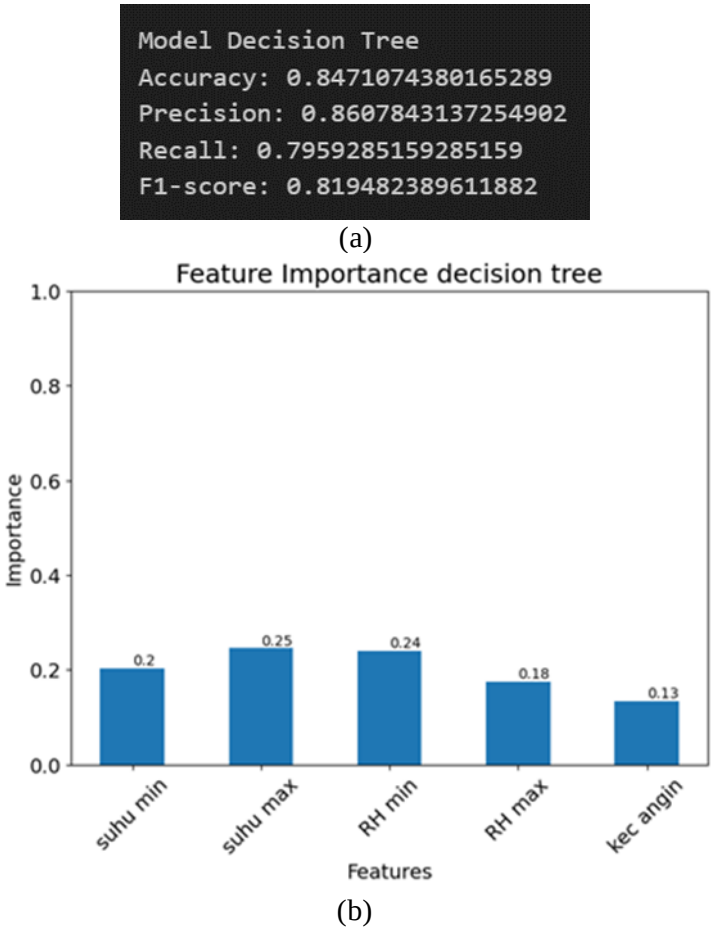
(a)



(b)

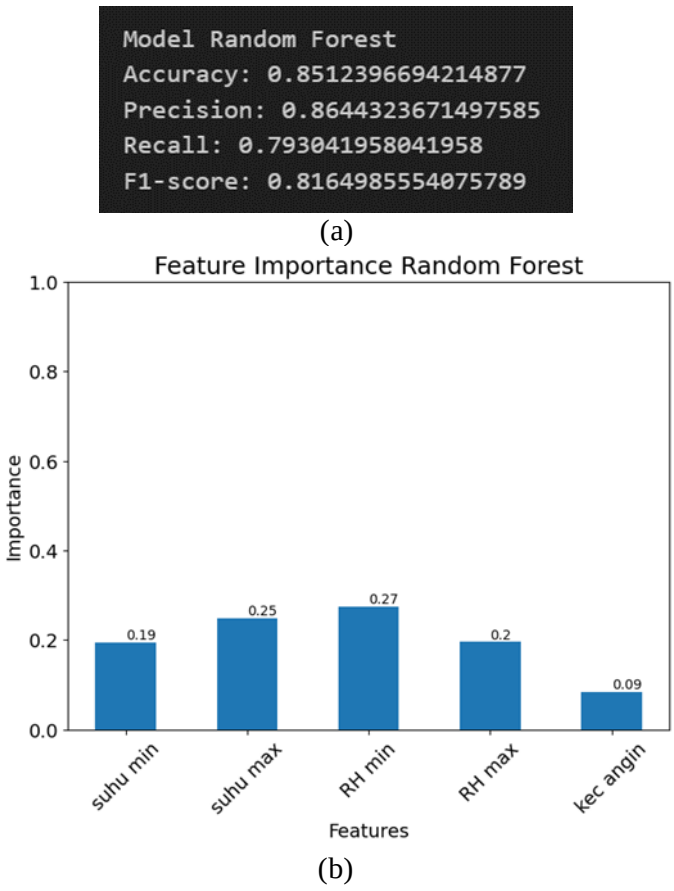
GAMBAR 5. *Output* Model *Gradient Boosting* (a) Hasil Metrik Evaluasi dan (b) Grafik *Feature Importance*

Pada model *gradient boosting* memiliki akurasi dan presisi ± 81% serta *recall* dan *F1-score* mendekati ±79%, dapat dikatakan bahwa model cocok untuk data tersebut. Dari plot *feature importance* diatas, dilihat bahwa fitur RH min dan RH max memiliki nilai sebesar ± 0.28, kedua fitur ini yang paling membantu model ini untuk menghasilkan prediksi cuaca. Fitur RH min dan RH memiliki nilai paling tinggi, kedua fitur ini saling berkorelasi satu sama lain. RH min dan RH max ini merupakan besar kelembaban udara maksimum dan minimum pada waktu itu. Pada model *gradient boosting* kedua fitur ini sangat berpengaruh terhadap model prediksi. Sedangkan untuk model *decision tree* memiliki hasil sebagai berikut:



GAMBAR 6. *Output* Model *Decision Tree* (a) Hasil Metrik Evaluasi dan (b) Grafik *Feature Importance*

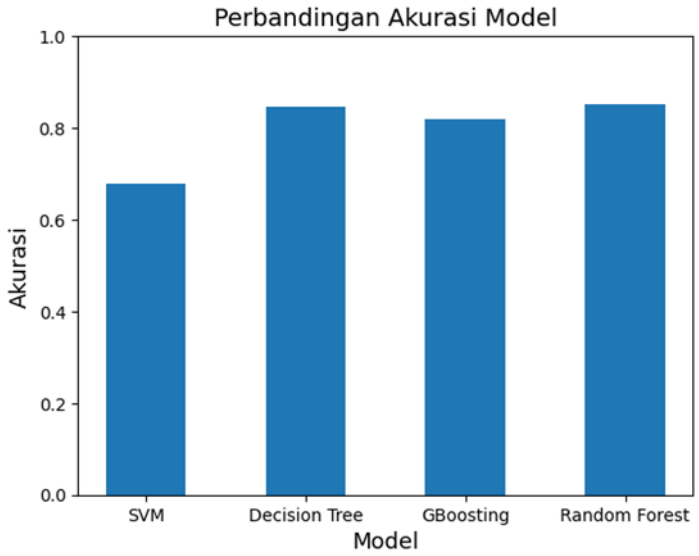
Pada hasil evaluasi metrik model *decision tree* memiliki akurasi sebesar 84.7% dengan presisi 86%. Model decision tree ini sudah memiliki akurasi yang cukup tinggi dan model dapat dikatakan cocok untuk data. Model ini memiliki hasil akurasi lebih tinggi daripada model sebelumnya yaitu model *gradient boosting*. Pada Gambar 7 terdapat hasil plot grafik *feature importance*, fitur suhu max dan RH min memiliki nilai paling tinggi sebesar ± 0.25. Hal ini menunjukan fitur yang paling berpengaruh dalam memprediksi cuaca adalah fitur suhu max dan RH (*Relative Humidity*) min, yaitu suhu maksimum udara dan kelembaban minimum udara yang diukur pada waktu setempat.



GAMBAR 7. *Output* Model *Random Forest* (a) Hasil Metrik Evaluasi dan (b) Grafik *Feature Importance*

Model *random forest* memiliki hasil akurasi dan presisi tertinggi diantara tiga model lainnya yaitu sebesar ± 86% dengan recall dan F1-score sebesar ± 81%, dengan nilai tersebut dapat dikatakan bahwa model cocok dengan data. Pada plot grafik *feature importance* diatas, terlihat bahwa fitur RH min memiliki nilai tertinggi sebesar ± 0.28. Hal ini membuktikan bahwa pada model *random forest*, fitur RH min berpengaruh lebih besar diantara empat fitur lainnya. RH min merupakan kelembaban minumim yang diukur pada waktu setempat.

Berikut merupakan hasil perbandingan akurasi model machine learning, pada grafik dapat terlihat bahwa model yang digunakan memiliki perbedaan hasil akurasi yang tipis. Pada model random forest dan decision tree akurasi yang dihasilkan tidak terlalu signifikan karena random forest memiliki akurasi 0. 8512396694214877 atau 85%, untuk decision tree memiliki akurasi 0.8471074380165289 atau 84% dengan selisih kedua model adalah 0.0042892562 atau ± 0.42%. dapat dikatakan model *random forest* dan *decision tree* sama baiknya dan cocok dengan data.



GAMBAR 8. *Output Plotting* Perbandingan Akurasi Model Machine Learning

SIMPULAN

Telah dilakukan pengujian data cuaca BMKG menggunakan model *support vector machine, gradient boosting, random forest, dan decision tree*. Hasil dari pengujian model *support vector machine* dengan akurasi ± 67%, untuk model *gradient boosting* dengan akurasi ± 81.8%, untuk model *random forest* dengan akurasi ± 85.1%, dan untuk model *decision tree* dengan akurasi ± 84.7%. Pada hasil pengujian akurasi model sudah cukup baik untuk prediksi cuaca karna 3 dari 4 model memiliki akurasi >80%. Namun, pada hasil pengujian ini model yang paling baik akurasi dan presisi-nya dalam model *random forest* dengan akurasi ±85% dan presisi ±86%.

REFERENSI

[1] D. Rahmalia, T. Herlambang, "Prediksi Cuaca Menggunakan Algoritma Particle Swarm Optimization-Neural Network (PSONN)," *Seminar Nasional Matematika dan Aplikasinya*, Surabaya, 2017.

[2] Indo, Intan *et al*., "Analisis Performansi Prakiraan Cuaca Menggunakan Algoritma Machine Learning," *Jurnal Pekommas*, vol. 6 no. 2, pp. 1-8, 2021.

[3] S. Chen *et al*., "Automated Poisoning Attacks and Defenses in Malware Detection Systems: An Adversarial Machine Learning Approach," 2017, [Online]. Available: http://arxiv.org/abs/1706.04146.

[4] F. Nelli, "Python data analytics: With Pandas, NumPy, and Matplotlib: Second edition," *Apress Media LLC*, 2018, doi: 10.1007/978-1-4842-3913-1.

[5] A. Natekin, A. Knoll, "Gradient boosting machines, a tutorial," *Front Neurorobot*, vol. 7, 2013, doi: 10.3389/fnbot.2013.00021.

[6] Y. Heryadi, T. Wahyono, "Machine Learning: Konsep dan Implementasi," Yogyakarta, 2020, [Online] Available: https://www.researchgate.net/publication/344419764.

[7] C. Bentéjac, A. Csörgő, G. Martínez-Muñoz, "A Comparative Analysis of XGBoost," 2019, doi: 10.1007/s10462-020-09896-5

[8] M. Raju, A. J. Laxmi, "IoT based Online Load Forecasting using Machine Learning Algorithms," *Procedia Computer Science*, pp. 552-560, 2020.

[9] Hemalatha *et al*., "Weather Prediction using Advanced Machine Learning Techniques," *Journal of Physics: Conference Series*, vol. 2089, no. 1, 2021.

[10] P. Fowdur et al., "A real-time collaborative machine learning based weather forecasting system with multiple predictor locations," Array 14, pp. 1-13, 2020.

[11] N. U. R. Ibrahim, T. F. Bacheramsyah, B. Hidayat, "Pengklasifikasian Grade Telur Ayam Negeri menggunakan Klasifikasi K-Nearest Neighbor berbasis Android," *Jurnal Teknik Energi Elektrik*, vol. 6, no. 2, pp. 288-302, 2018.

[12] L. A. Utami, "Analisis Sentimen Opini Publik Berita Kebakaran Hutan Melalui Komparasi Algoritma Support Vector Machine Dan K-Nearest Neighbor Berbasis Particle Swarm Optimization," *Jurnal Pilar Nusa Mandiri*, vol. 13, no. 1, pp. 103-112, 2017.

[13] M. Rangga, A. Nasution, M. Hayaty, "Perbandingan Akurasi dan Waktu Proses Algoritma K-NN dan SVM dalam Analisis Sentimen Twitter," *Jurnal Informatika*, vol. 6, no. 2, pp. 212-218, 2019, [Online]. Available: http://ejournal.bsi.ac.id/ejurnal/index.php/ji.