IT FOR SOCIETY, Vol. 04, No. 01
ISSN 2503-2224

Penggunaan Stacking Classifier Untuk Prediksi Curah Hujan

Diky Djafar Sidik
Departement of Computing
President Univeristy
Bekasi, Indonesia
diky.sidik@student.president.ac.id

DR. Tjong Wan Sen
Departement of Computing
President Univeristy
Bekasi, Indonesia
wansen@president.ac.id

***Abstract*—Curah hujan sebagai bentuk informasi dari data meteorologis, penting dalam segala kegiatan manusia yang berhubungan dengan alam, oleh karena itu prediksi atas curah hujan dengan hasil yang akurat merupakan hal yang sangat penting. Salah satu metode yang digunakan untuk prediksi/klasifikasi curah hujan adalah data mining dengan berbagai algoritma dan parameter data yang berbeda. Pada penelitian ini digunakan penggabungan metode klasifikasi dengan Teknik Ensemble Stacking/Stacked Generalization yang menggunakan Naïve Bayes dan C4.5 sebagai base learner dan KNN sebagai meta learner untuk klasifikasi curah hujan. Dataset yang dipergunakan adalah data klimatologi harian yang diambil dari website resmi BMKG (Badan Meteorologi, Klimatologi, Dan Geofisika) untuk stasiun UPT Bandung, Bogor, Citeko dan Jatiwangi dari periode 01 Januari 2000 sampai dengan 31 Desember 2018. Dengan menggunakan tiga skenario pengujian dan validasi menggunakan 10 fold cross validation diperoleh bahwa metode stacking dapat meningkatkan akurasi dari base classifier.**

Keywords—ensemble, stacking, naïve bayes, C4.5, KNN, curah hujan

I. PENDAHULUAN

Salah satu permasalahan yang paling menantang baik secara ilmiah dan juga teknologi di seluruh dunia saat ini adalah peramalan/prediksi kondisi cuaca dengan hasil seakurat mungkin [1], termasuk didalamnya adalah curah hujan, hal tersebut karena peramalan cuaca terkait dengan kepentingan aktivitas manusia dan disebabkan oportunisme yang tercipta oleh kemajuan teknologi yang terkait dengan bidang penelitian ini, seperti evolusi perhitungan dan peningkatan sistem pengukuran [2].

Seiring dengan meningkatnya ketersediaan data mengenai iklim memungkinkan *Data Mining* digunakan sebagai metode untuk memprediksi curah hujan, dengan berbagai algortima dan parameter data yang berbeda, beberapa metode dapat lebih akurat daripada metode yang lain [3].

Penelitian yang terkait dengan penggunaan data mining untuk prediksi/klasifikasi curah hujan telah banyak dilakukan oleh para peneliti dengan berbagai algoritma dengan parameter data yang berbeda diantaranya oleh : [4] mengenai curah hujan di kota Malang menggunakan *Naïve Bayes* mendapat hasil 97.74% akurasi, [5] melakukan klasifikasi curah hujan menggunakan algortima C4.5 di Kabupaten Bandung hasilnya adalah rata-rata nilai akurasi yang didapat 60% tanpa *pruning* dan 93.33% menggunakan *pruning*. [6] meneliti tentang estimasi curah hujan bulanan di Australia menggunakan metode *ensemble hybrid Genetic Algorithm* menghasilkan bahwa teknik *ensemble* menunjukan hasil akurasi terbaik.

[7] melakukan prakiraan cuaca menggunakan metode *Naïve Bayes*, dengan 4 *dataset*, hasilnya didapatkan untuk *dataset sample* diperoleh akurasi sebesar 81.66%, *dataset* Pune 93.52%, *dataset* Mumbai 90.69%, dan dataset Delhi 96.15%. [8] mengaplikasikan model *ensemble adaboost* pada SVM (adaSVM) dan *Naïve Bayes* (adaNaive) untuk menganalisa data curah hujan dengan periode waktu panjang didapatkan hasil AdaNaive dan AdaSVM lebih bagus hasilnya terhadap *dataset* terpilih dibandingkan SVM dan *Naïve Bayes*. [9] menggunakan algortima C4.5 untuk mendapatkan pola klasifikasi prediksi cuaca kedepan menghasilkan nilai akurasi sebesar 88.89%.

Hanya saja belum ada yang menggunakan penggabungan metode dengan teknik *ensemble stacking/stacked generalization* untuk prediksi/klasifikasi curah hujan pada penelitian sebelumnya sehingga hal tersebut yang mendasari penelitian ini. Pada penelitian ini digunakan penggabungan metode klasifikasi dengan teknik *ensemble stacking/stacked generalization* yang menggunakan *Naïve Bayes* dan C4.5 sebagai *base learner* dan KNN sebagai *meta learner* untuk klasifikasi curah hujan.

II. KAJIAN LITERATUR

A. Cuaca

Yang dimaksud dengan istilah cuaca adalah total dari seluruh variabel atmosfer pada tempat tertentu dalam periode waktu pendek, berbeda dengan iklim yang berbicara tentang perilaku atmosfer dalam jangka waktu yang lama dan merupakan kondisi yang sifatnya kumulatif dari keadaan rata-rata cuaca seperti pada angin, temperatur dan presipitasi.

Yang menjadi variabel utama dari cuaca adalah sinar matahari, kelembaban, awan, hujan dan angin serta tekanan atmosfer. Tekanan atmosfer dapat mempengaruhi kecepatan angin dan menentukan arah angin sehingga akan menggerakkan massa udara yang berbeda kelembaban dan

temperaturnya dari satu tempat ke tempat yang lainnya. Udara bergerak secara horizontal dan vertikal. Udara yang pergerakannya vertikal, berpotensi membentuk awan hujan [10].

B. Curah Hujan

Pengertian curah hujan menurut BMKG (Badan Meteorologi, Klimatologi dan Geofisika) adalah ketinggian air hujan yang terkumpul dalam penakar hujan pada tempat yang datar, tidak menyerap, tidak meresap dan tidak mengalir. Curah hujan 1 (satu) milimeter artinya dalam luasan satu meter persegi pada tempat yang datar tertampung air hujan setinggi satu milimeter atau tertampung air hujan sebanyak satu liter [11].

C. Faktor-faktor yang mempengaruhi curah hujan

Ada beberapa faktor yang dapat mempengaruhi jumlah curah hujan yang jatuh pada suatu tempat di Indonesia, menurut Sandy (2009) di [10] yaitu bergantung pada hal-hal berikut:

- Lokasi Daerah Konvergensi Antar Tropik (DKAT)
- Bentuk medan
- Arah angin
- Jarak perjalanan angin

Sedangkan menurut [12] faktor-faktor yang mempengaruhi curah hujan adalah kelembaban udara, tekanan udara, kecepatan angin dan suhu udara.

D. Naïve Bayes

Naïve Bayes merupakan suatu metode klasifikasi berdasarkan probabilistik sederhana. Algoritma ini menggunakan teorema Bayes yang mengasumsikan bahwa semua atribut independen diberi nilai oleh variabel kelas[13].

Naïve Bayes ditemukan oleh seorang ilmuwan Inggris yang bernama Thomas Bayes dan merupakan metode untuk memprediksi probabilitas kejadian di masa depan berdasarkan pengalaman sebelumnya. Salah satu keunggulan *Naive Bayes* adalah metode ini tidak membutuhkan data pelatihan yang banyak, dan hasil akurasinya cukup tinggi [4].

Dasar dari *Naive Bayes* adalah didasarkan pada penyederhanaan asumsi bahwa nilai suatu atribut adalah independen jika diberi nilai output.

Berikut ini adalah rumus dari Teorema Bayes :

$$P(H|X) = \frac{P(H)P(X|H)}{P(X)} \quad (1)$$

Dimana :

X = Data *sample* yang label kelasnya belum diketahui.

H = hipotesis data *sample* X

P(H) = probabilitas dari H *hypothesis* (probabilitas *prior*)

P(X) = Probabilitas dari *sample* data yang diamati

P(X|H) = probabilitas dari *sample* data X dengan asumsi hipotesis benar

Rumus di atas juga dapat ditulis sebagai berikut :

$$Posterior = \frac{prior\ x\ likelihood}{evidence} \quad (2)$$

E. C4.5

Algoritma C4.5 adalah salah satu contoh algoritma klasifikasi yang merupakan ekstensi Quinlan untuk algoritma ID3, C4.5 merupakan anggota dari *Decission Tree* yang dapat menangani data numerik dan diskret. *Decission tree* sendiri dikenal sebagai salah satu dari Teknik modeling yang sangat populer dan banyak digunakan pada klasifikasi maupun prediksi [13].

Decision tree dapat dipergunakan untuk mengeksplorasi data, dan menemukan link yang tersembunyi antara sejumlah variabel input dan variabel target. Tujuan dari decission tree adalah mengubah tabel data menjadi model pohon, dan kemudian menyederhanakan aturan [5].

Konsep *Entropy*

Nilai entropi digunakan sebagai parameter untuk mengukur heterogenitas sampel data. Semakin besar nilai entropi maka semakin heterogen, entropi ditentukan sebagai berikut :

$$Entropy\ (p) = -\sum_{i=1}^{n} P_i\ log_2\ P_i \quad (3)$$

Di mana, p adalah suatu set tertentu, n adalah jumlah partisi p, Pi adalah proporsi Pi ke p.

Gain Information

Setelah menemukan nilai entropi dari setiap atribut, kemudian dilanjutkan dengan mencari nilai derajat tertinggi dari pohon keputusan menggunakan informasi *gain* [13], didefinisikan sebagai berikut :

T didefinisikan sebagai *gain test* dari posisi p.

$$Gain\ (p) = Entropy\ (p) - info(p,T) \quad (4)$$

$$info\ (\mathrm{p,T}) = \sum_{i=1}^{n} \left(\frac{|Pi|}{|P|}\ x\ Entropy(Pi)\right) \quad (5)$$

Di mana nilai (PI) adalah semua nilai yang mungkin dari atribut T, nilai ini dapat menentukan urutan atribut dan membuat pohon keputusan dengan memetakan informasi gain terbesar.

F. KNN

KNN (K *Nearest Neighbor*) adalah sebuah *classifier nonparametric* yang *powerful*, yang mengklasifikasikan suatu pola yang belum terklasifikasi kedalam class yang mewakili berdasarkan dari mayoritas k tetangga terdekatnya [14].

K-*Nearest Neighbor* (KNN) merupakan contoh supervised learning dimana klasifikasi terhadap *instance* yang baru berdasarkan pada mayoritas dari kategori pada KNN. Algoritma ini bertujuan untuk mengklasifikasikan obyek baru berdasarkan pada atribut dan sample-sample dari hasil training.

Cara kerja dari Algoritma KNN cukup sederhana, yaitu bekerja berdasarkan jarak terpendek sebanyak k dari *query instance* ke training sample untuk menentukan KNN-nya. Sebagai contoh misal sebuah instance baru akan ditandai sebagai kelas A apabila kelas A merupakan klasifikasi yang paling banyak ditemui pada k buah tetangga terdekat dari instance tersebut. Untuk ukuran dekat dan jauhnya tetangga/*neighbor* biasanya dihitung berdasarkan *Euclidean Distance* sebagai berikut :

$$D(a,b) = \sqrt{\sum_{k=1}^{d}(a_k - b_k)^2} \qquad (6)$$

G. Metode Ensemble

Metode *ensemble* merupakan proses yang menggunakan beberapa macam model dasar untuk memprediksi suatu hasil, yang tujuannya adalah untuk mengurangi kesalahan prediksi, selama model dasar yang digunakan berbeda dan mandiri [15].

Menurut [16], *Ensemble classifier* adalah metode yang menggunakan atau menggabungkan beberapa pengklasifikasi untuk meningkatkan performance klasifikasi. Teknik ini lebih tahan terhadap noise dibandingkan dengan penggunaan *classifier* tunggal. Metode ini menggunakan pendekatan "*divide and conquer*" di mana masalah yang rumit diuraikan menjadi beberapa sub-masalah yang lebih mudah untuk dipahami dan diselesaikan. Contoh dari teknik *ensemble classifier* yaitu : *Bagging, Boosting dan Stacking*.

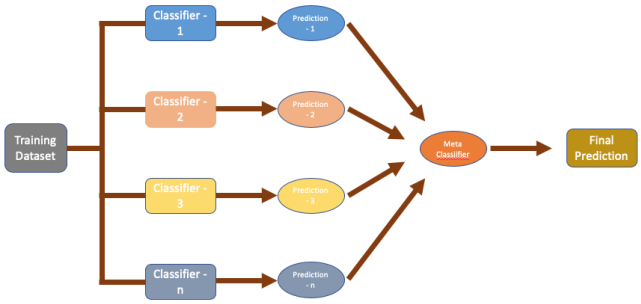
Keberhasilan teknik *ensemble* tergantung pada keragaman (*diversity*) dalam masing-masing individu *classifier* berkaitan dengan kesalahan klasifikasi [16]. Keragaman sangat krusial terhadap performa dari ensemble, hanya saja tidak mudah untuk menghasilkan keragaman ini dan masih belum adanya pemahaman yang jelas mengenai keragaman ini karena belum adanya definisi formal dari keragaman (*diversity*) itu sendiri [17].

H. Stacking Ensemble

Stacking adalah sebuah prosedur umum dimana sebuah *learner* ditraining untuk menggabungkan beberapa individual learner yang disebut *first-level learners*, sedangkan yang menggabungkan disebut *second-level learner* atau *meta-learner* [17].

Menurut [16] *Stacking* atau *Stacked generalization* adalah teknik lain untuk menggabungkan multi classifier. Tidak seperti teknik *bagging* dan *boosting*, *stacking* digunakan untuk menggabungkan *classifier* yang berbeda, misal : *decision tree, neural network, rule induction, naïve bayes, logistic regression* dan lain-lain.

Stacking terdiri dari dua level yaitu *base learner* sebagai *level-0* dan *stacking* model *learner* sebagai *level-1* atau *meta learner*. *Base learner* (level-0) menggunakan model yang berbeda untuk belajar dari suatu *dataset*. Output dari masing-masing model dikumpulkan untuk membuat dataset baru. Dalam dataset baru, setiap *instance* berhubungan dengan nilai sesungguhnya yang seharusnya diprediksi. Kemudian *dataset* tersebut digunakan oleh *stacking model learner* (level-1) untuk memberikan hasil akhir.



Gambar 1.Tahapan Metode *Stacking*

Algortima *Stacking* :

Input: Data set $D = \{(x_1,y_1), (x_2,y_2),..., (x_m,y_m)\}$;
First-level learning algorithms $L_1,... , L_T$;
Second-level learning algorithm L.

Process:
1. **for** t =1,...,T: % Training *learner* level 0
2. ht = Lt(D); % algoritma *learner* level 0
3. **end**
4. $D' = \emptyset$; % Membuat dataset baru
5. **for** i =1,...,m:
6. for t =1,... ,T:
7. $z_{it} = h_t(x_i)$;
8. **end**
9. D' = D' ∪ ((z_i1,...,z_iT),yi);
10. **end**
11. $h' = L(D')$; % Training *learner* level 1
% Algoritma *learner* level 1 pada dataset baru

Output: $H(x)= h'(h_1(x),... ,h_T(x))$

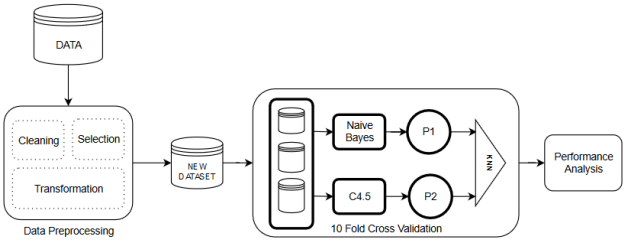
Langkah-langkah *Stacking* (Wolpert 1992)

1. Memisahkan data training menjadi dua bagian
2. Melakukan training pada beberapa *base learner* menggunakan bagian data 1
3. Membuat prediksi dengan *base learner* menggunakan bagian data 2
4. Menggunakan hasil prediksi dari langkah 3 sebagai *input* (data train) *learner* kedua

III. METODE PENELITIAN

Tujuan dari penelitian ini adalah untuk melihat performance dari metode *Stacking Classifier* dengan *base learner* menggunakan metode *Naïve Bayes* dan C4.5 dan KNN sebagai *meta learner* dalam prediksi/klasifikasi curah hujan, selain itu untuk mendapatkan gambaran pola pengaruh dari faktor-faktor cuaca terhadap curah hujan *Dataset* yang digunakan adalah data klimatologi harian yang diambil dari website resmi BMKG.

A. Alur Penelitian



Gambar 2.Tahapan / Alur Penelitian

Adapun alur dari penelitian ini adalah dimulai dari pengumpulan data iklim harian klmatologi dari website BMKG, kemudian setelah itu melalui proses *preprocessing* data berupa penggabungan file data, *cleaning missing /inconsistence* data, kemudian dilanjutkan ke diskretisasi data curah hujan. Selanjutnya terhadap data baru yang telah dilakukan proses preprocessing, dilakukan proses pembentukan model klasifikasi menggunakan metode *stacking* dan valdasi menggunakan *10 fold cross validation* dan terakhir dilakukan pengukuran *performance* menggunakan *confusion matrix*.

B. Dataset

Dataset yang digunakan pada penelitian ini adalah data iklim harian klimatologi hasil pengamatan unsur cuaca yang diunduh dari situs resmi BMKG Indonesia yaitu pada http://dataonline.bmkg.go.id/. Data yang dikumpulkan adalah data dari Stasiun Geofisika Bandung, Bogor, Citeko, Majalengka dengan periode 01 Januari 2000 sampai dengan 31 Desember 2018. Dataset memiliki 10 atribut dan 1 target label yaitu RR (curah hujan) :

1. Tanggal
2. Tavg = Temperatur rata-rata (°C)
3. Tn = Temperatur minimum (°C)
4. Tx = Temperatur maksimum (°C)
5. RH_avg = Kelembaban rata-rata (%)
6. Ss = Lamanya penyinaran matahari (jam)
7. ff_avg = Kecepatan angin rata-rata (m/s)
8. ff_x = Kecepatan angin maksimum (m/s)
9. ddd_x = Arah angin saat kecepatan maksimum (°)
10. ddd_car = Arah angin terbanyak (°)
11. RR = Curah hujan (mm)

Data hasil *download* masing-masing berupa file dengan format Excel, berisi data iklim harian perbulan sesuai dengan masing-masing daerah/kota.

C. Praprocessing Data

Sebelum *dataset* dapat digunakan perlu dilakukan tahap *praprocessing* terhadap data terlebih dahulu yaitu diantaranya adalah melakukan penggabungan file-file hasil download menjadi satu file *dataset* berdasarkan daerah, kemudian setelah itu dilanjutkan dengan melakukan pengecekan dan pembersihan pada *missing value*, *inconsistence data*, *duplicate data*, diskretisasi dan hal-hal yang perlu dilakukan agar dapat meningkatkan tingkat akurasi [13].

Tanggal	Tn	Tx	Tavg	RH_avg	RR	ss	ff_x	ddd_x	ff_avg	ddd_car
2013-09-01	20	29,8	24,4	72	0	6,5	2	270	1	W
2013-09-02	20,3	29,4	23,9	70	0	6,2	2	888	1	S
2013-09-03	19	28,4	23,1	72	0	8	3	225	2	SW
2013-09-04	19,5	29,2	23,9	65	0	6,4	3	225	2	SW
2013-09-05	17,7	29,7	23	70	0	5,4	4	180	1	SW
2013-09-06	17,8	29,5	23,5	68	1	5,6	2	45	1	SW
2013-09-07	18,8	30,4	23,3	68	0	7,7	3	45	2	NE
2013-09-08	18,8	30,2	23,4	72	0	6,4	3	45	2	NE
2013-09-09	20,3	30,6	24,9	68	0	7,8	4	45	2	W
2013-09-10	17,5	30,2	23,4	56	0	8	3	45	2	NE
2013-09-11	17,8	30	23,2	68	0	7	3	45	2	S
2013-09-12	18,6	30,8	23,6	66	0	5,4	3	180	2	S
2013-09-13					0	3,8	26	80	7	E
2013-09-14	18,5	30,4	23,4	70	34,4	5,7	4	45	2	SW
2013-09-15	19,7	30,4	24,7	68	8888	6,2	4	225	2	S
2013-09-16					0		1	240	0	N
2013-09-17							3	245	2	SW
2013-09-18	19,8	29,6	23,4	79	62	5,7	3	135	2	NE
2013-09-19					11	4,8	2	210	2	SW
2013-09-20	19,6	29,1	23,6	80	55,3	5	3	180	2	W
2013-09-21	19,6	29,8	23,7	76	7	6,1	2	135	1	NW
2013-09-22	17,6	30,2	23,7	66	1	6,2	3	45	1	NW
2013-09-23	18	30	23,9	63	0	6,5	3	270	2	W
2013-09-24	19,4	31,1	24,3	63	0	6,6	4	45	2	NE
2013-09-25	19,6	32,1	25	61	0	7	4	360	2	N
2013-09-26	19	30,2	24,5	71	0	5,8	2	315	2	N
2013-09-27	19,2	30,8	24,2	66	0	7,8	3	180	2	N
2013-09-28	19,8	30,1	24,1	70	8888	6,5	4	135	2	E
2013-09-29	19,8	30	23,7	70	0	5,4	4	45	1	NE
2013-09-30	19,8	28,3	23,2	79	8888	2,6	2	135	1	SE

Gambar 3. *Missing/Inconsistence Value*

Untuk mengklasifikasikan curah hujan pada penelitian ini dibuat sebuah atribut baru yaitu Kategori Hujan. Atribut baru ini merupakan hasil diskretisasi dari atribut Curah Hujan dengan cara memberi kategori pada variabel kategori hujan berdasarkan ketentuan press release BMKG tahun 2010. Dimana pada penelitian ini ada dua tipe pengkategorian / class kategori berdasarkan target 5 class dan kategori berdasarkan target 2 class, sehingga menjadi sebagai berikut :

Tabel 1. Kategori curah hujan target 5 Kelas

KATEGORI	KETERANGAN
Tidak Hujan	0 - 4 mm/hari
Ringan	1 - 5 mm/jam atau 5 - 20 mm/hari
Sedang	5-10 mm/jam atau 20 - 50 mm/hari
Lebat	10 - 20 mm/jam atau 50 - 100 mm/hari
Sangat Lebat	> 20 mm/jam atau > 100 mm/hari

Tabel 2. Kategori curah hujan target 2 Kelas

KATEGORI	KETERANGAN
Tidak Hujan	0 - 4 mm/hari
Hujan	>= 5 mm/hari

D. Proses Klasifikasi

Setelah data melalui proses *praprocessing* maka dilanjutkan dengan proses *modelling* klasifikasi. *Dataset* akan diklasifikasi baik menggunakan masing-masing algoritma klasifikasi *Naïve Bayes* dan C4.5 secara individu juga menggunakan metode *ensemble* menggunakan *stacking* yaitu penggabungan antara algoritma C4.5 dan Naive Bayes

sebagai *base classifier* kemudian hasil dari masing-masing algoritma *base classifier* akan dijadikan input kepada metode selanjutnya yang bertindak sebagai *meta classifier* dalam hal ini adalah KNN.

Untuk memvalidasi keakuratan sebuah model digunakan teknik *cross validation*, pada penelitian ini digunakan Teknik 10-*Fold Cross Validation* dimana data dibagi menjadi 10 bagian yang memiliki rasio yang sama / hampir sama.

Gambar 4. Proses 10 Fold Cross Validation

E. *Performance Measure*

Confusion Matrix merupakan tabel informasi mengenai perbandingan hasil klasifikasi antara aktual dan prediksi untuk setiap algoritma klasifikasi yang diberikan. Misal, untuk klasifikasi dengan jumlah data sebesar 1000 data dan memiliki 2 kelas ditunjukkan pada Tabel. Dalam tabel algoritma dapat mengklasifikasikan 450 data positif dengan benar, dan 430 data negatif dengan benar. Akan tetapi, salah klasifikasi sebanyak 70 data positif sebagai negatif, dan 50 data negatif sebagai positif.

Tabel 3. *Confusion Matrix*

Confusion Matrix		**Actual Class**	
		Positive	**Negative**
Predicted Class	**Positive**	**450**	**50**
	Negative	**70**	**430**

Kinerja algoritma dihitung berdasarkan angka-angka ini, seperti yang dijelaskan pada bagian berikut.

1. *Accuracy* : persentase dari jumlah prediksi yang benar dari seluruh jumlah prediksi yang dilakukan oleh *classifier*.

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} * 100\% \quad (7)$$

2. *Recall* : adalah persentase dari prediksi *True Positive*, dibandingkan dengan keseluruhan data *positive*.

$$recall = \frac{TP}{TP+FN} * 100\% \quad (8)$$

3. *Precision* : adalah ukuran persentase dari prediksi *True Positive* dibandingkan keseluruhan hasil yang diprediksi sebagai *positive*.

$$precision = \frac{TP}{TP+FP} * 100\% \quad (9)$$

4. *F-Measure* : adalah ukuran rata-rata dari precission dan recall

$$F1 - Score = \frac{2*(\text{precision}*recall)}{\text{precision+recall}} \quad (10)$$

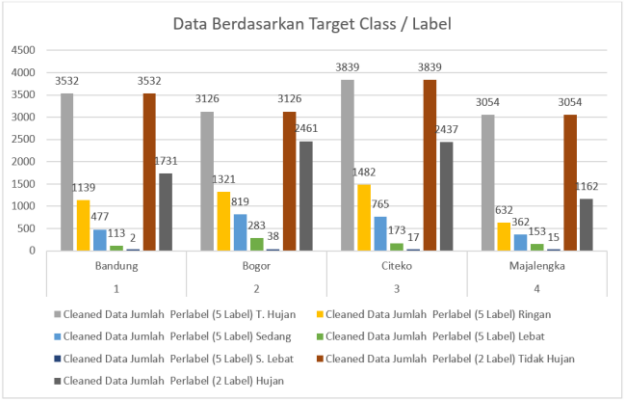
IV. HASIL PENELITIAN

Pada penelitian ini dilakukan proses klasifikasi/prediksi curah hujan dengan menggunakan algoritma *Stacking Classifier* dengan *Naïve Bayes*, C4.5 sebagai *base classifier* dan metode KNN sebagai *meta learner*, berdasarkan proses tersebut diharapkan dapat menggambarkan performance dari metode *Stacked Generalization/Stacking Classifier* dalam mengklasifikasi curah hujan dibandingkan dengan hasil yang didapatkan oleh base classifier.

A. *Skenario Proses Klasifikasi*

Dengan kondisi jumlah data pada masing-masing label target, baik dengan menggunakan 5 label target maupun yang menggunakan 2 label target adalah tidak seimbang, yaitu jumlah data dengan label Tidak Hujan yang mendominasi, dan data dengan label Sangat Lebat berjumlah sangat sedikit pada setiap dataset. Maka skenario pengujian dibagi menjadi tiga skenario :

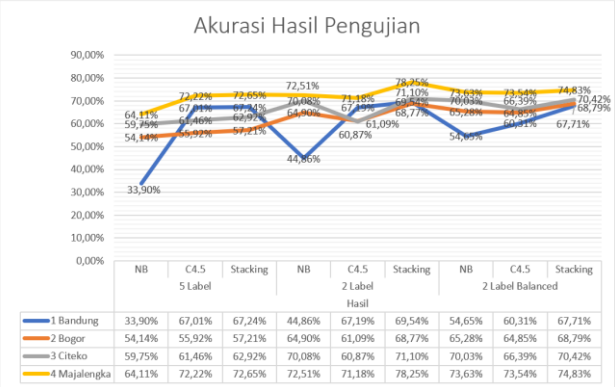
1. Klasifikasi masing-masing dataset menggunakan algoritma *Naïve Bayes*, C4.5 dan *Stacking* dengan jumlah label target 5 class.
2. Klasifikasi masing-masing dataset menggunakan algoritma *Naïve Bayes*, C4.5 dan *Stacking* dengan jumlah label target 2 class.
3. Klasifikasi masing-masing dataset menggunakan algoritma *Naïve Bayes*, C4.5 dan *Stacking* dengan jumlah label target 2 class yang sudah diseimbangkan.



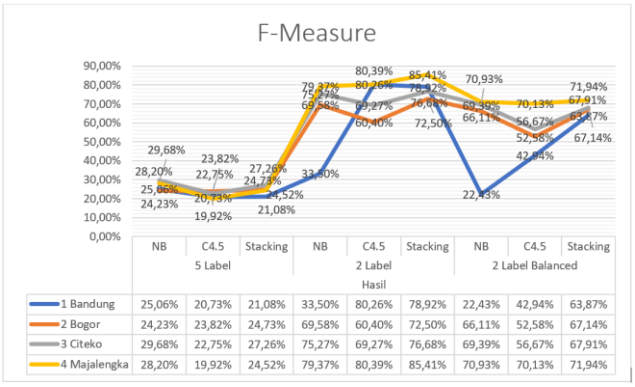
Gambar 5. Grafik Jumlah Data Per Target *Class* / Label

B. *Hasil Pengujian*

Dalam mengukur performance hasil dari proses modelling klasifikasi pada peneilitan ini digunakan ukuran akurasi dan F1-Measure sebagaimana dapat dilihat pada grafik di bawah ini :



Gambar 6. Grafik hasil akurasi pengujian



Gambar 7. Grafik hasil F-Measure pengujian

Berdasarkan grafik di atas dapat diambil beberapa point diantaranya :

- Nilai akurasi terendah adalah hasil klasifikasi menggunakan Naïve Bayes dengan target 5 class pada dataset Bandung yaitu sebesar 33,90%
- Nilai akurasi tertinggi terdapat pada dataset Majalengka yaitu sebesar 78,25% hasil klasifikasi menggunakan Stacking dengan target 2 class.
- Nilai F-1 Score terendah terdapat pada dataset Majalengka yaitu sebesar 19,92% hasil klasifikasi menggunakan C4.5 dengan target 5 class.
- Nilai F-1 Score tertinggi terdapat pada dataset Majalengka yaitu sebesar 85,41% hasil klasifikasi menggunakan Stacking dengan target 2 class.
- Performance dari masing-masing metode cenderung meningkat pada dataset dengan jumlah target class (label) lebih sedikit / jumlah data antar label seimbang.

V. KESIMPULAN

Penelitian ini melakukan pengujian terhadap metode ensemble stacking yang menggunakan *Naïve Bayes* dan C4.5 sebagai *base classifier* sedangkan KNN digunakan sebagai *meta learner* dalam klasifikasi curah hujan, menggunakan tiga skenario berbeda dan validasi menggunakan 10 *fold cross validation.*

Untuk mengukur performance hasil pengujian digunakan ukuran akurasi dan F1-Measure. Dari hasil pengujian didapatkan bahwa nilai akurasi tertinggi terdapat pada dataset Majalengka yaitu sebesar 78,25% hasil klasifikasi menggunakan Stacking dengan target 2 class begitupun untuk F1-Score tertinggi didapat pada *dataset* Majalengka yaitu sebesar 85,41% hasil klasifikasi menggunakan *Stacking* dengan target 2 *class.*

Berdasarkan hasil pengujian juga dapat dibuktikan bahwa metode *Stacking* berhasil meningkatkan performa dari *base classifier* pada semua skenario pengujian.

DAFTAR PUSTAKA

[1] J. Joseph and R. T K, “Rainfall Prediction using Data Mining Techniques,” *Int. J. Comput. Appl.*, vol. 83, no. 8, pp. 11–15, 2013.

[2] D. M. Casas, J. Ángel, T. González, J. Enrique, A. Rodríguez, and J. V. Pet, “Using Data-Mining for Short-Term Rainfall Forecasting,” pp. 487–488, 2009.

[3] E. G. Petre, “A Decision Tree for Weather Prediction,” *Bul. Univ. Pet. – Gaze din Ploieşti*, vol. LXI, no. 1, pp. 77–82, 2009.

[4] M. Muthmainnah, M. Ashar, I. M. Wirawan, and T. Widiyaningtyas, “Time Series Forecast for Rainfall Intensity in Malang City with Naive Bayes Methodology,” *3rd Int. Conf. Sustain. Inf. Eng. Technol. SIET 2018 - Proc.*, pp. 137–141, 2018.

[5] J. A. Suyatno, F. Nhita, and A. A. Rohmawati, “Rainfall forecasting in Bandung regency using C4.5 algorithm,” *2018 6th Int. Conf. Inf. Commun. Technol. ICoICT 2018*, vol. 0, no. c, pp. 324–328, 2018.

[6] A. Haidar, B. Verma, and T. Sinha, “A Novel Approach for Optimizing Ensemble Components in Rainfall Prediction,” *2018 IEEE Congr. Evol. Comput. CEC 2018 - Proc.*, no. 978, pp. 1–8, 2018.

[7] C. C. Janbandhu, P. D. Meshram, and M. N. Gedam, “Modelling Rainfall Prediction Using Data Mining Method - A Bayesian Approach,” *Int. J. Futur. Revolut. Comput. Sci. Commun. Eng.*, vol. 3, no. 11, pp. 472–474, 2017.

[8] B. Narayanan and M. Govindarajan, “Rainfall Prediction based on Ensemble Model,” pp. 8237–8243, 2016.

[9] I. Novandya, Adhika., Oktria, “Penerapan Algoritma Klasifikasi Data Mining C4.5 Pada Dataset Cuaca Wilayah Bekasi,” *J. Format*, vol. 6, no. 2, pp. 98–106, 2017.

[10] R. Pratama, “POLA CURAH HUJAN DI PULAU JAWA PADA PERIODE NORMAL, EL NINO DAN LA NINA,” 2011.

[11] BMKG, “Peraturan Kepala Badan Meteorologi, Klimatologi dan Geofisika,” no. 497, p. 4246703, 2010.

[12] N. Pradipta, P. Sembiring, and P. Bangun, "Analisis Pengaruh Curah Hujan Di Kota Medan," *Saintia Mat.*, vol. 1, no. 5, pp. 459–468, 2013.

[13] J. Han, M. Kamber, and J. Pei, *Data mining: Data mining concepts and techniques*, 3rd ed. Morgan Kaufmann Publishers, 2012.

[14] M. Huang, R. Lin, S. Huang, and T. Xing, "A novel approach for precipitation forecast via improved K-nearest neighbor algorithm," *Adv. Eng. Informatics*, vol. 33, pp. 89–95, 2017.

[15] V. Kotu and B. Deshpande, *Data Science Concepts and Practice*, Second. Morgan Kaufmann, 2019.

[16] I. Syarif, E. Zaluska, A. Prugel-Bennett, and G. Wills, "Application of bagging, boosting and stacking to intrusion detection," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7376 LNAI, pp. 593–602, 2012.

[17] Z.-H. Zhou, *Ensemble Methods Foundations and Algorithms*. CRC Press, 2012.

[18] Wirjohamidjojo, S., & Swarinoto, Y. S. (2013). Meteorologi Sinoptik. Pusat Penelitian dan Pengembangan Badan Meteorologi Klimatologi dan Geofisika Jl