Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 1, No.1, March 2018, pp. 27-31
p-ISSN:2614-3372 | e-ISSN: 2614-6150

Random Forest Algorithm for Prediction of Precipitation

[1]Aji Primajaya, [2]Betha Nurina Sari
[1,2]Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang
Email: [1]aji.primajaya@staff.unsika.ac.id, [2]betha.nurina@staff.unsika.ac.id

Article Info
Article history:
Received Feb 5th, 2018
Revised Feb 25th, 2018
Accepted Mar 10th, 2018

Keywords:
Precipitation
Prediction
Random Forest
Flood
Datamining

ABSTRACT
Predicting rainfall needs to be done as one of such effort to anticipate water flooding. One of the algorithm that can be used to predict rainfall is random forest. The porpose of the research is to create a model by implementing random forest algorithm. The research method consist of four steps: data collection, data processing, random forest implementation, analysis. Random forest implementation with using training set resulted model that has accurracy 71,09%, precision 0.75, recall 0.85, f-measure 0.79, kappa statistic 0.33, MAE 0.35, RMSE 0.46, ROC Area 0.78. Implementation of random forest algorithm with 10-fold cross validation resulted the output with accurracy 99.45%, precision 0.99, recall 0.99, f-measure 0.99, kappa statistic 0.99, MAE 0,09, RMSE 0.14, ROC area 1.

Corresponding Author:
Aji Primajaya
Teknik Informatika, Fakultas Ilmu Komputer
Universitas Singaperbangsa Karawang
Jl. HS. Ronggowaluyo, Kel.Sirnabaya, Kec.Telukjambe Timur, Kab.Karawang, 41361
Email: aji.primajaya@staff.unsika.ac.id

1. PENDAHULUAN

Curah hujan yang tinggi akan bisa memperbesar potensi terjadinya banjir. Penelitian yang dilakukan Nugroho pada tahun 2008 memberikan informasi bahwa tingginya curah hujan telah menyebabkan banjir di Jakarta pada tahun 2002, tepatnya antara bulan Januari sampai dengan bulan Februari [1]. Prediksi curah hujan merupakan salah satu cara yang bisa digunakan untuk mengantisipasi terjadinya banjir sehingga curah hujan akan mempengaruhi tindakan yang bisa dilakukan untuk mengantisipasi terjadinya banjir [1]. Jika pemerintah atau *stakeholder* mengetahui tinggi rendahnya curah hujan maka akan bisa memberikan dukungan terkait upaya untuk pencegahan dan penanganan banjir. Oleh sebab itu, studi ini membuat model untuk prediksi curah hujan.

Penelitian terkait prediksi curah hujan telah berhasil dilakukan dengan menggunakan berbagai metode, seperti menggunakan metode: (1) C4.5 [2][3]; (2) *random forest* [4]; dan (4) *classification and regression trees* (CART) [5]. Penelitian ini menggunakan metode *random forest* untuk memprediksi curah hujan. Random forest adalah salahsatu metode berbasis klasifikasi dan regresi dimana terdapat proses agregasi pohon keputusan [5]. Metode ini dipilih karena menghasilkan kesalahan yang lebih rendah, memberikan akurasi yang bagus dalam klasifikasi, dapat menangani data pelatihan yang jumlahnya sangat besar, dan efektif untuk mengatasi data yang tidak lengkap [6].

Prediksi curah hujan dengan menggunakan *random forest* sebenarnya sudah pernah dilakukan oleh peneliti sebelumnya. Dhawangkhara dan Riksakomara pada tahun 2017 menggunakan algoritma *random forest* untuk melakukan prediksi terkait curah hujan [5]. Penelitian tersebut hanya menggunakan 5 buah parameter, yaitu: (1) suhu udara; (2) titik embun; (3) kecepatan angina; (4) tekanan udara; dan (5) visibilitas. Sedangkan penelitian ini menggunakan lima buah parameter yang pernah digunakan Raditya pada tahun 2012 untuk pola prediksi hujan menggunakan algoritma C4.5 [7], yaitu: (1) suhu rata-rata (TEMP*)*; (2) *mean sea level pressure* (SLP); (3) *mean station pressure* (STP); (4) kecepatan angin rata-rata (WDSP); dan (5) kecepatan angin maksimum (MXSPD). Meskipun menggunakan algoritma C4.5, penelitian tersebut [7] berhasil menghasilkan akurasi yang lebih baik daripada penelitian yang menggunakan algoritma *random forest* [5]. Oleh sebab itu, penelitian ini mencoba menggunakan atribut yang digunakan Raditya [5] untuk meningkatkan akurasi prediksi curah hujan menggunakan algoritma *random forest*.

2. METODOLOGI PENELITIAN

Penelitian prediksi curah hujan dengan menggunakan algoritma Random Forest menggunakan metode penelitian berjumlah 4 tahapan. Tahapan tersebut adalah: (1) pengumpulan data; (2) tahap pengolahan data; (3) tahap implementasi random forest; dan (4) tahap analisis. Metode penelitian yang dilakukan pada penelitian ini dapat dilihat pada Gambar 1.

Gambar 1. *Flowchart* metode penelitian

2.1. Pengumpulan Data

Data dari penelitian ini diambil di https://www1.ncdc.noaa.gov/pub/orders/ pada bulan Juli 2017. Data tersebut berasal dari stasiun Achmad Yani pada 1 Mei 2010 sampai dengan 4 Juni 2010. Dataset terdiri dari 2188 data dengan 16 atribut, yaitu: (1) *station* (STN); (2) *weather bureau air force navy* (WBAN); (3) tahun bulan dan hari (YEARMODA); (4) TEMP; (5) rata-rata dew point dalam sehari (DEWP); (6) SLP; (7) STP; (8) rata-rata visibility per hari (VISIB); (9) WDSP; (10) MXSPD; (11) maksimum GUST; (12) maksimum suhu (MAX), (13) minimum suhu (MIN), (14) PRCP, (15) kedalaman salju (SNDP); dan (16) indikator cuaca (FRSHTT).

2.2. Pengolahan Data

Pada tahap pengolahan data dilakukan pemilihan atribut yang digunakan, pembersihan data, transformasi data. Atribut dipilih berdasarkan hasil penelitian Raditya pada tahun 2012 [7], yaitu: (1) TEMP; (2) SLP; (3) STP; (4) WDSP; dan (5) MXSPD. Kelas dari klasifikasi ini diambil dari atribut PRCP. Pembersihan data dilakukan dengan menghapus data kosong yang ditemukan pada atribut MXSPD dan PRCP. Jumlah data kosong yang ditemukan adalah 205 data. Hasil dari pembersihan data didapatkan 1983 data yang akan diolah untuk tahap selanjutnya. Redundansi, inkonsistensi dan *outlier* tidak ditemukan. Fase pengolahan data selanjutnya adalah transformasi data. Tipe data dari atribut PRCP yang awalnya berupa real dengan ditambah kode pengukuran curah hujan, diubah terlebih dahulu ke dalam tipe numerikdengan menghapus kode tersebut. Selanjutnya format data numerikal diubah menjadi tipe data kategorikal, yaitu hujan atau tidak hujan. Apabila PRCP bernilai 0 maka tidak hujan, sebaliknya jika ada nilai PRCP lebih dari 0 maka Hujan. Contoh transformasi data yang dilakukan pada atribut PRCP dapat dilihat pada Tabel 1.

Tabel1. Transformasi data PRCP

PRCP dengan Kode	PRCP tanpa Kode	PRCP
0.02A	0.02	Hujan
0.02A.	0.02	Hujan
0.00I	0.00	Tidak
0.00I	0.00	Tidak
0.79B	0.79	Hujan
0.00I	0.00	Tidak
0.16B	0.16	Hujan
0.00I	0.00	Tidak
0.00I	0.00	Tidak
2.01B	2.01	Hujan

Hasil dari tahap pengolahan data ini adalah dataset yang siap untuk diproses pada algoritma *random forest*. Keterangan jenis atribut dari dataset yang siap diaolah dapat dilihat pada Tabel 2.

Tabel2. Jenis atribut data

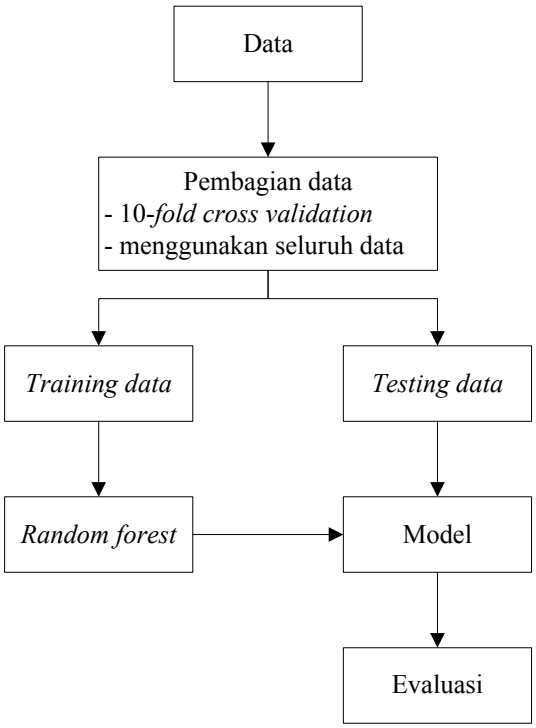
No.	Nama Atribut	Tipe data
1.	Suhu rata-rata (TEMP)	Real
2.	*Mean Sea Level Pressure* (SLP)	Real
3.	*Mean Station Pressure* (STP)	Real
4.	kecepatan angin rata-rata (WDSP)	Real
5.	kecepatan angin maksimum (MXSPD)	Real
6.	Precipitation (PRCP)	Kategorikal

2.3. Implementasi *Random Forest*

Random forest merupakan salah satu metode yang digunakan untuk klasifikasi dan regresi. Metode ini merupakan sebuah *ensemble* (kumpulan) metode pembelajaran menggunakan pohon keputusan sebagai

base classifier yang dibangun dan dikombinasikan [8]. Ada tiga aspek penting dalam metode *random forest*, yaitu: (1) melakukan *bootstrap sampling* untuk membangun pohon prediksi; (2) masing-masing pohon keputusan memprediksi dengan prediktor acak; (3) lalu *random forest* melakukan prediksi dengan mengombinasikan hasil dari setiap pohon keputusan dengan cara *majority vote* untuk klasifikasi atau rata-rata untuk regresi [9].

Pada Gambar 3 diberikan informasi terkait dengan langkahimplementasi algoritma *random forest* untuk prediksi curah hujan. Langkah awal adalah melakukan input data hasil dari transformasi data dimana terdiri dari atribut penjelas dan atribut target. Setalah itu data dibagi menjadi dua jenis (*training data* dan *testing data*) dengan menggunakan metode K-Fold Cross Validation dimana nilai K dipilih 10. Selain itu, penentuan *training data* dan *testing data* juga dilakukan dengan menggunakan seluruh data. Nantinya akan dilakukan perbandingan hasil antara kedua jenis metode penentuan *training data* dan *testing data* tersebut. Algoritma *random forest* pada penelitian ini menggunakan sepuluh pohon keputusan yang dibangkitkan secara acak. *Training data* digunakan sebagai data masukan untuk algoritma *random forest* sedangkan *testing data* digunakan untuk menguji atau mengevaluasi output atau model yang dihasilkan dari algoritma *random forest*.



Gambar 3. Implementasi *random forest*

Evaluasi performa random forest dilakukan dengan menggunakan beberapa parameter pengukuran, yaitu akurasi, presisi, *recall*, *f-measure*. Akurasi merupakan parameter yang paling umum dan sederhana untuk mengevaluasi performa algoritma klasifikasi, yaitu dengan menunjukkan berapa tingkat atau presentase kebenaran prediksi. Presisi, *recall* dan *f-measure* adalah parameteryang sering digunakan dalam *information retrieval*, dimana presisi adalah tingkat ketepatan hasil prediksi, *recall* adalah tingkat sensitivitas terhadap bagian data yang relevan, sedangkan *f-measure* adalah rata-rata harmoni dari presisi dan *recall*.

Selain empat parameter pengukuran performa klasifikasi, evaluasi performa algoritma *random forest* bisa juga dengan *kappa statistic*, *mean absolute error* (MAE), *root mean squared error* (RMSE), dan *receiver operating characteristic* (ROC) Area [10]. *Kappa statistic* digunakan untuk mengukur performa algoritma klasifikasi atau bisa mengestimasi kemiripan antar anggota dari *ensemble* dalam *multi-classifier systems*. MAE menunjukkan berapa banyak penyimpangan prediksi dari yang sebenarnya. RMSE disebut sebagai *brier score* yang mengukur terkait penyimpangan prediksi dari yang sebenarnya. ROC Area menunjukkan metrik seberapa baik model dalam memprediksi. ROC dapat digunakan untuk menganalisis model hasil klasifikasi [11].

2.4. Analisis

Pada tahap analisis dilakukan analisis terhadap model yang dihasilkan dalam hubungannya dengan studi kasus prediksi curah hujan. Selain itu, hasil *testing* berdasarkan *parameter testing* juga dianalisis untuk mengetahui kualitas dari model yang dihasilkan.

3. HASIL DAN ANALISIS

Tabel 2 merupakan penyajian evaluasi output dari algoritma *random forest* dengan teknik pembagian data menggunakan 10-*fold cross validation*. Tabel 3 merupakan hasil evaluasi dari model yang dihasilkan oleh algoritma *random forest* dengan cara menggunakan seluruh data sebagai *training data* dan *testing data*.

Table 2. Evaluasi performa *random forest* pada hasil model dari 10-*fold cross validation*

No	Parameter	Nilai
1.	Akurasi	71,6087 %
2.	Presisi	0,748
3.	*Recall*	0,854
4.	F-Measure	0,798
5.	Kappa statistic	0,3286
6.	MAE	0,351
7.	RMSE	0,4609
8.	ROC Area	0,778

Table 3. Evaluasi performa *random forest* pada hasil model dari penggunaan seluruh data

No	Parameter	Nilai
1.	Akurasi	99,4453 %
2.	Presisi	0,994
3.	*Recall*	0,998
4.	F-Measure	0,996
5.	Kappa statistic	0,9877
6.	MAE	0,0864
7.	RMSE	0,1408
8.	ROC Area	1

Berdasarkan hasil evaluasi model yang dihasilkan maka dapat dianalisis bahwa implementasi *random forest* lebih cocok menggunakan seluruh data untuk *training data* dan *testing data* jika dilihat dari evaluasi akurasi, presisi, *recall*, *f-measure*, *kappa statistic*, MAE, RMSE dan ROC *Area*. Akurasi *random forest* dengan menggunakan teknik 10-*fold cross validation* sebesar 71,09% sedangkan dengan teknik menggunakan seluruh data sebesar 99,45%. Tingkat akurasi yang dihasilkan dari penggunaan teknik seluruh data sebagai *training data* dan *testing data* merupakan *resubtitution estimate*, dimana hasil perkiraannya sering sangat bagus yang berguna untuk tujuan diagnostik.

4. KESIMPULAN

Berdasarkan hasil dan analisis terkait implementasi *random forest*, maka bisa disimpulkan bahwa *random forest* mampu digunakan untuk menghasilkan model prediksi hujan. Penggunaan teknik penentuan *training data* dan *testing data* menggunakan kesuluruhan data mampu menghasilkan model yang lebih baik dibandingkan dengan teknik pembagian data k-*fold cross validation*.

UCAPAN TERIMAKASIH

Kami ucapkan trimakasih kepada sumber data yang telah menyediakan datanya dengan gratis sehingga bisa digunakan pada penelitian ini.

REFERENSI

[1] S. P. Nugroho. "Analisis Curah Hujan Penyebab Banjir Besar di Jakarta Pada Awal Februari 2007". *JAI*. 2008; 4(1): 50-55.
[2] A. Novandya, I. Oktria. "Penerapan Algoritma Klasifikasi Datamining C4.5 Pada Dataset Cuaca Wilayah Bekasi". *Jurnal Format*. 2017; 6(2): 98-106.
[3] A. Khusaeri, S. Ilham, D. Nurhasanah, D. Delpidat, Anggri, A. Primajaya, B. N. Sari. "Algoritma C4.5 untuk Pemodelan Daerah Rawan Banjir Studi Kasus Daerah Karawang Jawa Barat". *ILKOM Jurnal Ilmiah*. 2017; 9(2): 132-136
[4] Y. Wang, S. Shia, Q. Tang, J. Wu, X. Zhu. "A Novel Consistent Random Forest Framework: Bernoulli Random Forest". *IEEE Transaction On Neural Network and Learning Systems.* 2017; 1-14

[5] M. Dhawangkhara, E. Riksakomara. "Prediksi Intensitas Hujan Kota Surabaya dengan Matlab Menggunakan Teknik Random Forest dan CART (Studi Kasus Kota Surabaya)". *Jurnal Teknik ITS.* 2017; 6(1): 94-99
[6] L. Breiman. Random Forests. Machine Learning. 2001; 4(1): 5-32
[7] A. Raditya. "*Implementasi Datamining Classification untuk Mencari pola Prediksi Hujan dengan Menggunakan Algoritma C4.5*". Dalam: Publikasi Jurusan Teknik Informatika, Fakultas Teknologi Industri, Universitas Gunadarma. 2012.
[8] V. Y. Kulkarni, P. K. Sinha. "Effective Learning and Classification Using Random Forest Algorithm". *International Journal of Engineering and Innovative Technology (IJEIT)*. 2014; 3(11): 267-273
[9] M. G. Sadewo, A. P. Windarto, D. Hartama D. "Penerapan Datamining Pada Populasi Daging Ayam RAS Pedaging di Indonesia Berdasarkan Provinsi Menggunakan K-Means Clustering". *InfoTekJar (Jurnal Nasional Informatika dan Teknik Jaraingan)*. 2017; 2(1): 60-67
[10] C. Ferri, J. Hernández-Orallo, R. Modroiu. "An Experimental Comparison of Performance Measures for Classification". *Pattern Recognitipns Letters.* 2009; 30(1): 27-38
[11] S. Dewi. "Komparasi 5 Metode Algoritma Klasifikasi Datamining Pada Prediksi Keberhasilan Pemasaran Produk Layanan Perbankan". *Jurnal Techno Nusa Mandiri*. 2016; 13(1): 60-65.

BIBLIOGRAFI PENULIS

Aji Primajaya, S.Si., M.Kom, lahir di Pacitan pada 26 April 2987. Saat ini penulis adalah dosen Teknik Informatika, Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang. Lulus S1 dari jurusan Fisika, Universitas Negeri Surabaya pada tahun 2010, lalu melanjutkan studi S2 Ilmu Komputer di Institut Pertanian Bogor, lulus pada tahun 2015. Penulis tergabung dalam grup riset data mining di Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang.

Betha Nurina Sari, M.Kom, lahir di Kediri pada 23 Oktober 1989. Lulus S1 dari program studi Ilmu Komputer, Universitas Brawijaya tahun 2012. Pada tahun 2013, penulis melanjutkan studi S2 Ilmu Komputer di Universitas Indonesia dan lulus pada tahun 2015. Saat ini penulis adalah dosen Teknik Informatika, Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang dan tergabung dalam grup riset data mining.