Jurnal TAM (Technology Acceptance Model) Volume 9, Nomor 2, Desember 2018 Hal. 74-78
p-ISSN : 2339-1103
e-ISSN : 2579-4221

PERBANDINGAN ALGORITMA CART DAN *K-NEAREST NEIGHBOR* UNTUK PREDIKSI LUAS LAHAN PANEN TANAMAN PADI DI KABUPATEN KARAWANG

Muhammad Fadhlil Aziz[1], Sofi Defiyanti[2], Betha Nurina Sari[3]
[1]Prodi Teknik Informatika Universitas Singaperbangsa Karawang
[2]Prodi Teknik Informatika Universitas Singaperbangsa Karawang
[3]Prodi Teknik Informatika Universitas Singaperbangsa Karawang
[1,2,3] Jl. H. S. Ronggowaluyo Telukjambe Timur Karawang 41361
Email : 1441177004260@student.unsika.ac.id[1], sofi.defiyanti@staff.unsika.ac.id[2], betha.nurina@staff.unsika.ac.id[3].

ABSTRAK
Kabupaten Karawang dikenal sebagai salah satu lumbung padi nasional karena terdapat banyak area pesawahan khususnya tanaman padi. Namun alih fungsi dari lahan pertanian menjadi area industri atau perumahan dapat merubah struktur geografis Kabupaten Karawang yang sebelumnya dipenuhi lahan pertanian menjadi area industri dan property. *Data mining* merupakan suatu teknik penggalian suatu informasi dari data yang berukuran besar. Salah satunya teknik regresi. Dalam memprediksi sesuatu dataset yang bertipe data numerik biasanya menggunakan teknik regresi. Pada penelitian ini digunakan teknik regresi untuk memprediksi luas lahan panen di Kabupaten Karawang dengan menggunakan tools WEKA 3.8.2. Perbandingan yang dihasilkan dilihat dari *correlation coefficient, mean absolute error,* dan *root mean squared error*. Pada perbandingan algoritma digunakan skenario yang sama yaitu *cross validation* 10 *folds*. Hasil uji coba dengan menggunakan skenario yang sama menunjukkan bahwa kedua algortima dapat digunakan untul memprediksi luas lahan panen di Kabupaten Karawang. Kesimpulan dari penelitian ini menunjukkan bahwa algoritma CART memiliki performa lebih baik dari algoritma KNN dengan correlation coefficient 0,9646, MAE 498,6229, dan RMSE 834,0204.

Kata kunci : CART, *data mining*, *k-nearest neighbo*,, luas lahan panen.

ABSTRACT
Karawang regency is known as one of the nation rice granaries because the are many areas of rice fields, especially rice. But the transfer of function from agricultural land into industrial or recidential area can change the geographical structure of Karawang regency previously filled with agricultural land into industrial and property areas. Data mining is a technique of extracting an information from large data. One of them regression techniques. In predicting something a dataset of a numeric data type usually uses a regression technique. In this study used regression techniques to predict the area of harvested land in Karawang regency by using tools WEKA 3.8.2. The resulting comparison is seen from correlation coefficient, mean absolute error, and root mean squared error. In comparison algorithm used the same scenario is cross validation 10 folds. The result of the experiment using the same scenario shows that both algorithm can be used to predict the area of harvest area in Karawang regency. The result of evalution with same scenario shows that CART algorithm has better performance than KNN algorithm with correlation coefficient 0,9646, MAE 498,6229, and RMSE 834,0204.

Keywords : area of harvest land, cart, data mining, k-nearest neighbor.

I. PENDAHULUAN

Kabupaten Karawang dikenal sebagai salah satu lumbung padi nasional karena terdapat banyak area pesawahan khususnya tanaman padi. Serta memiliki potensi yang cukup menjanjikan baik dari segi industri, properti atau segi yang lainnya. Namun alih fungsi dari lahan pertanian menjadi area industri atau perumahan dapat mengubah struktur geografis Kabupaten Karawang yang sebelumnya dipenuhi lahan pertanian menjadi area industri dan properti. Hal ini dapat menjadi penyebab berkurangnya luas lahan pertanian khususnya lahan pertanian tanaman padi di Kabupaten Karawang.

Dapat diketahui bahwa luas lahan panen tanaman padi di 30 kecamatan mengalami perubahan dari tahun ke tahun, ada yang mengalami kenaikan dan penurunan. Dari keenam tahun dapat dilihat bahwa luas lahan panen di Kabupaten Karawang yang paling rendah pada tahun 2015.

Data mining adalah proses yang menggunakan teknik statistic, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstrasi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar [1].

Penelitian ini diharapkan dapat memberikan rekomendasi algoritma yang dapat untuk menangani jenis data class target numerik atau regresi khususnya prediksi, dikarenakan algoritma yang direkomendasikan sudah melewati seleksi pengujian dengan teknik klasifikasi serta teknik regresi.

Pada penelitian Nataraharja [2], hasil evaluasi perbandingan algoritma C4.5 dan CART (*Classification and Regression Tree*) untuk memprediksi luas lahan panen tanaman padi di Kabupaten Karawang bahwa algoritma CART memiliki nilai akurasi, *precision*, *recall*, *f-measure*, dan *roc area* lebih tinggi dibandingkan dengan algoritma C4.5. Sedangkan pada penelitian Saraswati [3], hasil evaluasi perbandingan algoritma *naive bayes* dan *k-nearest neighbor* untuk prediksi luas lahan panen tanaman padi di Kabupaten Karawang bahwa algoritma *k-nearest neighbor* memiliki nilai akurasi, *recall*, dan *f-measure* lebih tinggi dibandingkan dengan algoritma *naive bayes*.

II. TINJAUAN PUSTAKA

2.1. *Data Mining*

Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan didalam *database*. *Data mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar [4].

Data mining membahas perihal penggalian atau pengumpulan informasi yang berguna dari kumpulan data. Informasi yang biasanya dikumpulkan adalah pola – pola tersembunyi pada data, hubungan antar elemen – elemen data, ataupun pembuatan model untuk keperluan peramalan data [5].

2.2. Regresi

Dalam machine learning, analisis regresi berusaha untuk memperkirakan hubungan antara variabel output dan satu set independen variabel input dengan secara otomatis belajar dari sejumlah akurasi sampel. Tujuan utama penerapan analisis regresi biasanya untuk mendapatkan prediksi yang tepat dari tingkat variabel output untuk sampel baru.

Contoh dari metode untuk analisis regresi dalam literatur adalah *linear regression*, *automated learning of algebraic models for optimisation* (ALAMO), *support vector regression* (SVR), *multilayer perception* (MLP), *K-nearest neighbor* (KNN), *multivariate adaptive regression splines* (MARS), dan *regression tree* [6].

2.3. Prediksi

Prediksi adalah memperkirakan sesuatu yang terjadi pada masa yang akan datang. Prediksi juga dapat digunakan dalam pengklasifikasian, tidak hanya untuk memprediksi time series, karena sifatnya yang bisa menghasilkan *class* berdasarkan atribut yang ada [7].

2.4. CART (*Classification And Regression Trees*)

CART (*Classification And Regresion Tree*) adalah metode statistik non parametrik yang digunakan untuk melakukan analisis klasifikasi. CART pertama kali diperkenalkan pada tahun 1984 oleh empat ilmuwan Amerika serikat yaitu Leo Breiman, Jerome H. Friedman, Richard A.Olshen, dan Charles J. Stone. CART terdiri dari dua analisis yaitu *classification trees* dan *regression trees*. Jika variabel yang dimiliki bertipe kategorik maka CART menghasilkan pohon klasifikasi (*classification trees*). Sedangkan jika variabel dependen yang dimiliki bertipe kontinu atau numerik maka CART menghasilkan pohon regresi (*regression trees*) [8].

2.5. *K-Nearest Neighbor* (KNN)

K-Nearest Neighbor (KNN) termasuk kelompok *instance-based learning*. Algoritma ini juga merupakan salah satu teknik *lazy learning*. KNN dilakukan dengan mencari kelompok k objek dalam *data training* yang paling dekat (mirip) dengan objek pada data baru atau data *testing* [9]. Algoritma *K-nearest Neighbor* adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. *Nearest Neighbor* adalah pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus baru dan kasus lama yaitu berdasarkan pada pencocokan bobot dari sejumlah fitur yang ada [10].

2.6. Evaluasi Model

Evaluasi model merupakan bagian integral dari proses pengembangan model. Ini membantu untuk menemukan model terbaik yang mewakili data kami dan seberapa baik model yang dipilih akan bekerja di masa depan. Mengevaluasi kinerja model dengan data yang digunakan untuk pelatihan tidak dapat diterima dalam ilmu data karena dapat dengan mudah menghasilkan model yang terlalu optimis dan berlebihan. Ada dua metode evaluasi model dalam ilmu data, *Hold-Out* dan *Cross-Validation*. Untuk menghindari *overfitting*, kedua metode menggunakan satu set tes (tidak dilihat oleh model) untuk mengevaluasi kinerja model [11].

1. *Hold-out*
   Dengan model ini, sebuah *dataset* akan diolah secara acak dibagi menjadi tiga bagian :
   a) *Training set* adalah bagian dari dataset yang digunakan untuk membuat model prediksi.
   b) *Validation set* adalah bagian dari dataset yang digunakan untuk menilai kinerja model yang dibangun di fase pelatihan. Ini menyediakan *platform* uji untuk parameter model *fine tuning* dan memilih model yang

berkinerja terbaik. Tidak semua algoritma pemodelan membutuhkan satu set validasi.

c) *Test set* atau contoh yang tidak terlihat adalah bagian dari dataset untuk menilai kemungkinan kinerja masa depan model. Jika suatu model sesuai dengan set pelatihan jauh lebih baik daripada yang cocok dengan set tes, *overfitting* mungkin adalah penyebabnya.

2. *Cross-validation*
Ketika hanya terdapat sejumlah data yang terbatas, untuk mencapai perkiraan yang tidak bias dari kinerja model, maka digunakan k-*fold cross-validation*. Dalam k-*fold cross-validation*, kami membagi data ke dalam himpunan bagian dengan ukuran yang sama. Kami membangun model k *folds*, setiap *folds* meninggalkan salah satu himpunan bagian dari pelatihan dan menggunakannya sebagai perangkat tes. Jika k sama dengan ukuran sampel, ini disebut "*leave-one-out*" [11].

2.7. WEKA 3.8.2

WEKA adalah sebuah paket *tools machine learning* praktis. WEKA merupakan singkatan dari *Waikato Environment for Knowledge Analysis*, yang dibuat di Universitas Waikato, New Zealand untuk penelitian, pendidikan dan berbagai aplikasi. WEKA mampu menyelesaikan masalah-masalah *data mining* di dunia nyata, khususnya klasifikasi yang mendasari pendekatan-pendekatan *machine learning*. Perangkat lunak ini ditulis dalam hirarki *class* Java dengan metode berorientasi objek dan dapat berjalan hampir di semua *platform* [12].

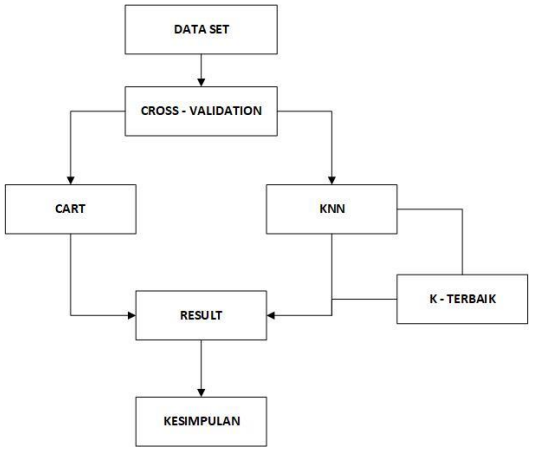
2.8. Padi

Padi (beras) merupakan bahan makan utama masyarakat Indonesia yang mencapai 255,46 juta orang dengan laju pertumbuhan sebesar 1,31% dan tingkat konsumsi beras mencapai 124,89 kg/kapita/tahun.

Prediksi permintaan padi untuk konsumsi pada tahun 2016 berdasarkan angka prognosa konsumsi beras perkapita tahun 2015 ditetapkan sebesar 124,89 kilogram/kapita/tahun. Dengan jumlah penduduk mencapai 258,71 juta orang maka diperkirakan kebutuhan beras untuk konsumsi langsung rakyat Indonesia mencapai 32,31 juta ton. [13].

III. METODOLOGI PENELITIAN

Penelitian ini menggunakan *dataset* luas lahan panen tanaman padi. *Dataset* tersebut nantinya akan diprediksi dengan menggunakan algoritma CART dan KNN dengan menggunakkan *test options* 10 *folds cross validation* dan metodologi KDD (*knowledge discovery in databases*) [14]. Skenario penelitian dapat dilihat pada Gambar 1.



Gambar 1. Skenario Penelitian

3.1. Dataset

Dataset yang digunakan bersumber dari Dinas Pertanian Kabupaten Karawang yang jumlah barisnya sebanyak 210 dengan 14 variabel antara lain : curah hujan, hari hujan, luas baku sawah, luas tanam, luas sawah, luas lahan panen, produksi, produktivitas, wbc, tikus, hpp, penggerek batang, siput murbai, bhd, dan blasit.

3.2. Pengujian Algoritma

Metode pengujian yang dilakukan pada penelitian ini adalah *cross validation* dengan 10 *folds* yang dilakukan terhadap algoritma *classification and regression tree* dan *k-nearest neighbor* untuk mengetahui algoritma mana yang memiliki performa yang lebih baik dengan melihat hasil dari *correlation coefficient*, *mean absolute error*, dan *root mean squared error* [11].

1. *Correlation Coefficient* akan dihitung dengan persamaan sebagai berikut :

$$r = \frac{n(\sum pa)-(\sum p)(\sum a)}{\sqrt{[n\sum p^2-(\sum p)^2][n\sum a^2-(\sum a)^2]}}$$

Keterangan :
a = *actual target*
p = *predicted target*
n = banyaknya data

2. *Mean absolute error* akan dihitung dengan persamaan sebagai berikut :

$$MAE = \frac{\sum_{i=1}^{n}|p_i-a_i|}{n}$$

Keterangan :
a = *actual target*
p = *predicted target*
n = banyaknya data

3. *Root mean squared error* akan dihitung dengan persamaan sebagai berikut :

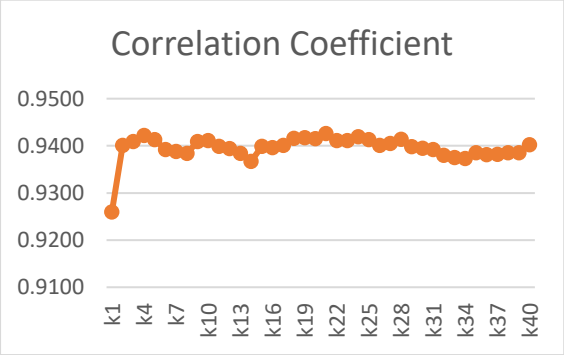
$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i-a_i)^2}{n}}$$

Keterangan :
a = *actual target*
p = *predicted target*
n = banyaknya data

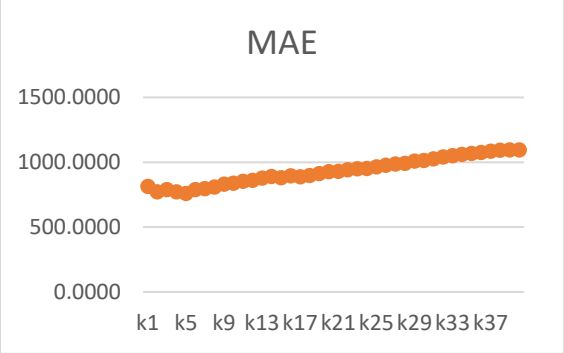
IV. HASIL PENELITIAN

4.1. Algoritma *K-Nearest Neighbor*

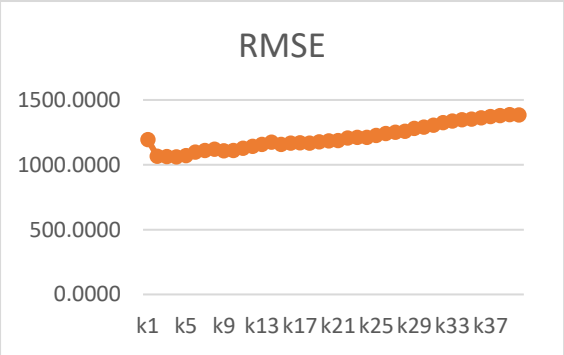
Pada algoritma KNN dilakukan percobaan k sebanyak 40 kali dengan menggunakan test options 10 *folds cross validation* yang bertujuan untuk menemukan *trend* yang nantinya akan digunakan untuk mengetahui informasi apakah dengan menggunakan nilai k berukuran besar akan menghasilkan sebuah nilai evaluasi yang bagus atau tidak. Hasil *trend* dengan melihat nilai ketiga teknik evaluasi dapat dilihat pada Gambar 2.



Gambar 2. *Trend Correlation Coefficient Algoritma* KNN



Gambar 3. *Trend Mean Absolute Error* Algoritma KNN



Gambar 4. *Trend Root Mean Squared Error* Algoritma KNN

Dari hasil percobaan sebanyak 40 kali bahwa semakin besar nilai k tidak menjadikan nilai evaluasi nya semakin baik, dan diambil rentang nilai untuk dijadikan nilai pembanding dengan nilai hasil dari algoritma CART. Dimana dari keseluruhan hasil percobaan didapatkan rentang nilai setiap teknik evaluasi sebagai berikut :

Tabel 1. Rentang Nilai Percobaan Algoritma KNN

Correlation coefficient	MAE	RMSE
0,926 s/d 0,9426	792,9733 s/d 1096,341	1060,545 s/d 1387,028

4.2. Algoritma *Classification and Regression Tree*

Untuk algoritma CART hanya dilakukan satu kali pengujian dengan menggunakan *cross – validation folds* 10 tanpa melakukan pruning dan mendapatkan hasil sebagai berikut:

Tabel 2. Hasil Percobaan Algoritma CART

Correlation coefficient	MAE	RMSE
0,9646	498,6229	834,0204

Berdasarkan hasil evaluasi dapat diketahui bahwa algoritma CART memiliki performa lebih baik dari algoritma KNN yang dapat dilihat pada hasil *correlation coefficient*, *mean absolute error*, dan *root mean squared error*. CART dikatakan lebih baik karena algoritma ini memiliki keunggulan yang tidak dimiliki oleh algoritma KNN yaitu, CART lebih mudah untuk diinterpretasikan, lebih akurat dan lebih cepat dalam perhitungannya, selain itu CART juga bisa menangani himpunan data besar [15]. Banyaknya himpunan data yang menyebabkan algoritma CART memiliki performa lebih baik dari algoritma KNN, pada proses perhitungan algoritma CART lebih cepat dibandingkan dengan algoritma KNN karena algoritma KNN pada proses perhitungannya harus menghitung jarak antar baris data berulang kali untuk menemukan rangking yang menghasilkan nilai prediksi sedangkan algoritma CART lebih mudah untuk diinterpretasi karena menghasilkan sebuah pohon keputusan yang dijadikan sebagai pola pengambil keputusan.

V. KESIMPULAN

Kesimpulan yang diperolah pada penelitian ini adalah dari hasil perbandingan algoritma *classification and regression tree* dan *k-nearest neighbor* untuk prediksi luas lahan panen tanaman padi di Kabupaten Karawang menyatakan bahwa algoritma *classification and regression tree* dengan *test option cross validation 10 folds* memiliki performa lebih baik dari algoritma *k-nearest neighbor* dilihat dari hasil *correlation coefficient* sebesar 0,9646 serta dinyatakan sebagai korelasi

sempurna, *mean absolute error* sebesar 498,6229, dan *root mean squared error* sebesar 834,0204.

REFERENSI

[1] Suprapto, "Penerapan data mining untuk memprediksi mahasiswa *drop out* menggunakan *support vector machine*," *Komputaki,* pp. 14-49, 2015.

[2] N. A. Nataraharja, "Perbandingan algoritma C4.5 dan algoritma CART untuk prediksi luas lahan panen tanaman padi di karawang," *Skripsi,* 2017.

[3] V. Saraswati, "Perbandingan algoritma naive bayes dan *k-nearest neighbor* untuk prediksi luas lahan panen tanaman padi di karawang," *Skripsi,* pp. 9-40, 2017.

[4] D. Kartika dan Pane, "Implementasi data mining pada penjualan produk elektronik dengan algoritma apriori (studi kasus : kreditplus)," *Pelita informatika budi darma,* pp. 25-29, 2013.

[5] S. Adinugroho dan Y. A. Sari, Implementasi data mining menggunakan WEKA, Malang: UB Press, 2018.

[6] L. Yang, S. Liu, S. Tsoka dan L. G. Papageorgiou, "*A regression tree approach using mathematical programming*," *ELSEVIER,* p. 347–357, 2017.

[7] V. Andriyana dan Y. S. Nugroho, "Perbandingan 3 metode dalam data mining untuk prediksi penerimaan beasiswa berdasarkan prestasi di SMA Negeri 6 Surakarta," pp. 1-8, 2015.

[8] A. Waluyo, M. A. Mukid dan T. Wuryandari, "Perbandingan klasifikasi nasabah kredit menggunakan regresi logistik biner dan CART (*classification and regression trees*)," *Media statistika,* pp. 95-104, 2014.

[9] H. Leidiyana, "Penerapan algoritma k-nearest neighbor untuk penentuan resiko kredit kepemilikan kendaraan bermotor," *Penelitian ilmu komputer,* pp. 65-76, 2013.

[10] R. I. Ndaumanu, Kursini dan M. R. Arief, "Analisa prediksi tingkat pengunduran diri mahasiswa dengan metode *k-nearest neighbor*," *Jatisi,* pp. 1-15, 2014.

[11] S. Sayad, *Real Time Data Mining The Future Is Here,* Toronto: ResearchGate, 2011.

[12] S. Pujiono, A. Amborowati dan M. Suyanto, "Analisis kepuasan publik menggunakan weka dalam mewujudkan *good governance* di kota yogyakarta," *DASI,* p. 4, 2013.

[13] Suwandi, Outlook komoditas pertanian sub sektor tanaman pangan (padi), Jakarta: Pusat data dan sistem informasi pertanian kementrian pertanian, 2016.

[14] J. Han, M. Kamber dan J. Pei, *Data mining concepts and techniques third edition*, Waltham: Elsevier, 2012.

[15] F. E. Pratiwi dan I. Zain, "Klasifikasi Pengangguran Terbuka Menggunakan CART (*Classification and Regression Tree*) di Provinsi Sulawesi Utara," *Sains dan Seni Promits,* pp. 2337-3520, 2014.