ANALISIS CART (CLASSIFICATION AND REGRESSION TREES) UNTUK PREDIKSI PENGGUNA SEPEDA BERDASARKAN CUACA

Annida Purnamawati[1)], Monikka Nur Winnarto[2)], Mely Mailasari[3)]

[1,3]Sistem Informasi, Universitas Bina Sarana Informatika
[2] Teknologi Informasi, Universitas Bina Sarana Informatika

[1]Jl. Ringroad Barat, Ambarketawang, Gamping, Sleman, Yogyakarta
[2,3] Jl. Kramat Raya No. 98, Senen, Jakarta Pusat

Email: [1]annida.npr@bsi.ac.id, [2]monikka.mnt@bsi.ac.id, [3]mely.myl@bsi.ac.id

Abstract

This study presents a rule-based classification model for user predictions based on weather. Cyclists are very popular because of the increased comfort and environment. The data used is public data from the Bike Sharing Dataset taken from Kaggle. The data has data on bicycle users every hour. With this data set, the authors managed to find the accuracy of the CART method which explains the accuracy of 96%. The results of this study show the estimated distribution under various bicycles with spatial variables, distribution of use and time including the most influential variables in the predictions of bicycle users.

Keyword: CART, *decission tree, bike, cuaca, transportasi*

Abstrak

Penelitian ini menyajikan model klasifikasi berbasis aturan untuk prediksi pengguna sepea berdasarkan cuaca. Pengguna sepeda sangat populer karena kenyamanan dan kelestarian lingkungan menjadi meningkat. Data yang digunakan merupakan data publik *dari Bike Sharing* Dataset yang diambil dari Kaggle. Data tersebut memiliki data pengguna sepeda setiap jam. Dengan dataset tersebut penulis berhasil menemukan akurasi dari metode CART yang menjelaskan hasil akurasinya mencapai 96%. Hasil dari penelitian tersebut dengan menunjukkan arsitektur perkiraan distribusi di bawah berbagai sepeda dengan spatiotemporal variable, pendistribusian penggunaan dan waktu termasuk di merupakan variabl yang paling berpengaruh dalam prediksi pengguna sepeda tersebut.

Kata Kunci: CART, *decission tree, sepeda, cuaca, transportasi*

1. Pendahuluan

Dalam kurun waktu beberapa dekade terdapat sistem untuk pengguna sepeda. Ini merupakan sistem transportasi yang di kembangkan untuk menyediakan sepeda untuk pengguna umum. Sistem sepeada ini memungkinkan untuk menyewa sepeda dari satu lokasi kelokasi lain, di mana pengguna dapat naik dan kemudian kembali ke lokasi yang telah di tentukan.

Motif utama mengenai sistem tersebut yaitu untuk memfokuskan pada lingkungan dan kesejahteraan sosial. Dengan kemajuan teknologi yang pesat dari sistem transportasi cerdas dan teknologi setelah tahun 2000-an, sistem sepeda ini digunakan secara global. Kemudian banyaknya pengguna sepeda juga di tentukan oleh perubahan iklim yang cukup signifikan yang akan menyebabkan cuaca dapat berubah sangat cepat serta sulit untuk di prediksi [1]. Dari perubahan cuaca yang sangat sulit di prediksi ini menjadikan jumlah masyarakat yang bersepeda juga mengalami peningkatan dan penurunan yang berubah-ubah jumlahnya [2].

Banyak Negara memiliki sistem ini, seperti Ddareungi yang merupakan sistem sepeda dari Korea Selatan dimulai dari tahun 2015 yang dikenal dengan Seoul Bik dalam bahasa Inggris [3]. Hal tersebut untuk mengatasi masalah seperti kenaikan harga minyak, kemacetan lalu lintas dan pencemaran lingkungan serta membangun lingkungan yang sehat.

Penelitian terdahulu dari Satishkumar, V. E dan Yongyun Co mengenai penelitiannya dengan lima model statistik yang dilatih serta dioptimalkan menggunakan beberapa pendekatan validasi silang berulang dan set pengujian menggunakan (a) CUBIST (b) Regularized Random Forest (c) Klasifikasi serta Pohon Regresi (d) K Tetangga Terdekat (e) Pohon Inferensi Bersyarat. Indeks yang mengevaluasi beberapa seperti R2, Root Mean Squared Kesalahan Mean Absolute Kesalahan dan Koefisien Variasi yang saat itu digunakan untuk mengukur prediksi perfor Mance dari model regresi. Dari penelitian yang menunjukkan bahwa model berbasis aturan CUBIST telah mampu menjelaskan sekitar 95% dan 89% Varians (R2) pada perangkat pengujian data Seoul Bike dan data

program *Capital Bike Share* berdasarkan penelitian tersebut [3].

Bike Sharing berbasis darat ini merupakan contoh eksplorasi eksperimen penelitian yang kemungkinan akan mendapatkan manfaat dari banyak multivarian teknik analisis data berkembang dalam beberapa tahun terakhir [4]. Secara singkat beberapa masalah yang timbul dalam analisis datanya, dan karakteristik umum dari cuaca termasuk di dalamnya suhu dan jam merupakan variabl yang paling berpengaruh dalam prediksi pengguna sepeda. Dalam studi kasus ini, berbagai macam teknik penerapan algoritma klasifikasi di terapkan untuk menentukan teknik mana yang tampaknya akurasi paling baik deskriminasi [5].

Dalam pendekatannya penulis, mendekatkan dengan karakteristik umum dari sinyal dan peristiwa latar belakang sekelompok node yang dekat antara satu sama lain pada tree tersebut yang di sesuaikan secara bersamaan. Karenanya pilihan akhir juga dipengaruhi oleh interaksi antara prediktor. Sebuah pohon tidak lunak bersama dengan data latih yang digunakan untuk konstruksinya. Sebagai masukan ke fase pengoptimalan, yang mencoba menemukan nilai parameter ambang lunak, yang menghasilkan perkiraan terbaik dari penelitian data pelatihan. Karena struktur pohon, khususnya jumlah nodenya, tetapi selama pengoptimalan tidak terjadinya overviting.

Tujuan dari pemrosesan pasca trees adalah untuk mencapai fungsi yang lebih baik yang lebih cocok dengan data pelatihan. Karena kompleksitas pengklasifikasian tidak meningkat terlalu banyak, orang mungkin berharap untuk mencapai kesalahan generalisasi yang lebih kecil. Selain pendekatan yang lebih baik di dalam kasus, dimana fungsi probabilitas bersyarat yang tiak di ketahui bersifat kontinu, penulis dapat memperoleh yang lebih baik ari perkiraan bahkan jika sebenarnya dari konfisional probabilitas membuat lompoatan pada batas antara wilayah dengan klasifikasi berbeda, jika batasnya tidak dalam arah sumbu-paralel. CART mungkin mewawkili fungsi yang lebih kompleks dari yang lain. Secara khusus sebagai konsekuensi dari interaksi beberapa prediktor fungsi preiksi dari tree mungin memiliki vektor gradien ke arah umum, sambil mempertahankan sejumlah kecil simpul dari original tree.

2. Tinjauan Pustaka

2.1 Data Mining

Data mining merupakan proses untuk mencari dan menemukan pola dengan menggunakan data yang jumlahnya besar, dengan bertujuan untuk menemukan pola yang sebelumnya tidak diketahui [6]. Data mining didefinisikan sebagai proses menemukan pola dalam data. Prosesnya harus otomatis atau (lebih sering) semiotomatis. Pola yang ditemukan harus bermakna karena mengarah pada beberapa keuntungan. Data selalu hadir dalam jumlah yang substansial [8]. Pada data mining kumpulan data pada masa lalu atau yang terkumpul dapat dijadikan sebagai sesuatu yang dapat diolah sehingga menghasilkan pengetahuan yang sangat berharga.

2.2 Metode LinearDiscriminantAnalysis (LDA)

Metode LDA dapat digunakan sebagai penentuan hubungan kebergantungan antara satu variable respon dengan dua atau lebih variable yang bebas dan berbentuk kategori yang sudah dikelompokkan dalam masing masing kelas. Metode ini tujuannya untuk mengklasifikasikan pengamatan ke dalam kelas dan hasilnya bersifat homogen sebagai pegamatan, selain itu metode ini juga bersifat heterogen antar kelas [9].

Metode LDA dapat melakukan analisis pada data serta dokumen yang berukuran besar. LDA menggunakan metode bag of words untuk mengidentifikasi informasi topik tersembunyi dalam kumpulan dokumen besar. Metode LDA menjadikan setiap dokumen sebagai vektor jumlah kata dan mewakili distribusi probabilitas beberapa topik, dimana setiap topik direpresentasikan sebagai distribusi probabilitas beberapa kata. Mekanisme kerja LDA dibagi menjadi dua bagian yaitu penalaran dan realisasi. Inferensi adalah proses metode LDA yang digunakan untuk menentukan bobot setiap kata dalam setiap dokumen dalam korpus. Implementasi merupakan tahapan dimana aplikasi LDA selanjutnya memenuhi kebutuhan temu kembali informasi [10].

2.3 Metode k-Nearest Neighbor (k-NN)

Metode k-NN merupakan salah satu metode klasifikasi objek dengan menggunakan data pembelajaran, dimana diambil jarak terdekat dengan objek tersebut. Data yang digunakan dapat di proyekskan kedalam banyak dimensi yang masin-masing dimensi merepresentasikan fitur yang diperoleh dari data [11].

K-Nearest Neighbor (K-NN) merupakan salah satu algoritma yang digunakan untuk menyelesaikan masalah pengklasifikasian. Prinsip kerja K-NN yaitu mencari jarak terdekat antar data yang akan dievaluasi dengan tetangga terdekat dalam data pelatihan. Algoritma ini sering menghasilkan hasil yang kompetitif dan signifikan. Untuk menghitung jarak menggunakan jarak Euclidean. Rumus jarak Euclidean didefinisikan dalam Persamaan (1) [12].

$$\boldsymbol{d_i} = \sqrt{\sum_{i=1}^{P}(\boldsymbol{X_{2i}} - \boldsymbol{X_{1i}})^2} \quad (1)$$

Keterangan :
X_1 : Data Latih
X_2 : Data Uji
i : Variabel Data
d : Jarak
p : Dimensi Data

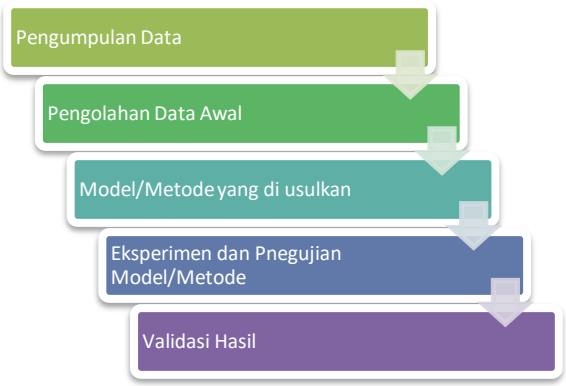
2.4 Metode Clasification and Regression Trees (CART)

Classification and Regression Trees (CART) merupakan suatu algoritma dari suatu teknik pohon keputusan dimaksud juga sebagai Decission Tree. CART disebut juga sebagai algoritma statistic juga nonparametric yang dapat digambarkan sebagai variable respon dengan satu atau bisa lebih variable predictor atau variable independent [13]. Variabel respon yang dimaksud dapam penelitian ini yaitu variable yang berskala

kategorik dimana dimaksudkan metode tersebut akan digunakan sebuah metode klasifikasi pohon.

3. Pembahasan

Pada penelitian ini data yang digunakan adalah Bike Share Daily Dataset yang diperoleh dari Kaggle dengan alamat web: https://www.kaggle.com/contactprad/bike-share-daily-data.
Dalam penelitian ini akan dilakukan beberapa langkah-langkah atau tahapan penelitian seperti gambah di bawah ini:



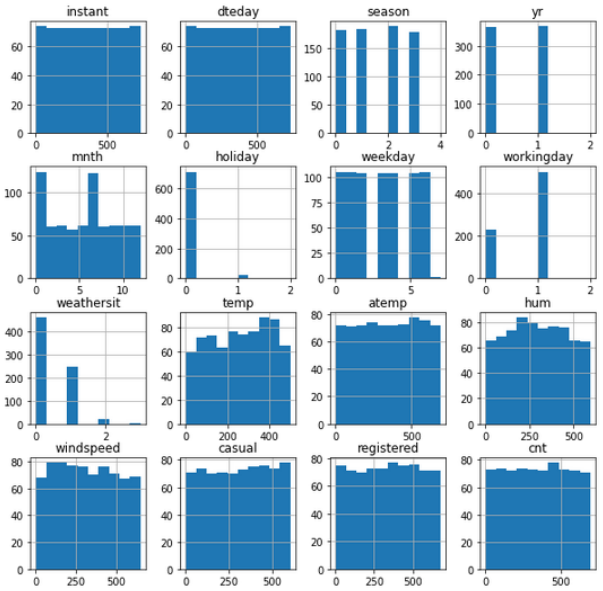
Gambar 1. Tahapan Penelitian

3.1 Pengumpulan Data

Penulis menggunakan teknik pengumpulan data dengan menggunakan pendekatan kuantitatif yaitu dengan melakukan survei dari literatur akademis dengan bidang keilmuan yang sama guna untuk mendapatkan konsep yang relevan dengan menyalurkan kajian inovasi penelitian dari kebijakan publik [14]. Dalam pengumpulan data ada 2 sumbr dengan data sekunder dan data primer yaitu diperoleh dengan key information sedangkan data sekunder diperoleh dari hasil analisis penulis terhadap jawaban key informations dan narasumber yang telah dikaitkan dengan tabel koding dan teori strategi dari public relations (dokumen dan data jurnal, buku ilmiah dan internet) [15].

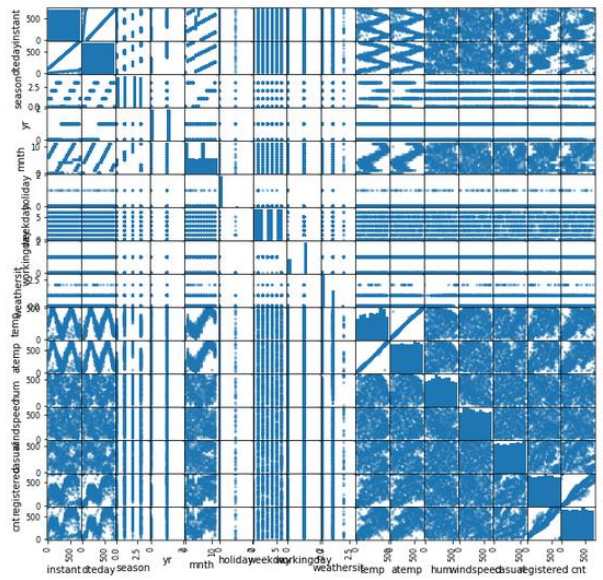
Pada penelitian ini termasuk dalam data sekunder karena data diperoleh dari Kaggle dengan judul Bike Share Daily Data (https://www.kaggle.com/contactprad/bike-share-daily-data).

Dari dataset tersebut di gambarkan pada Histogram berikut ini.



Gambar 2. Histogram Dataset

Tampilan *Scater Matrix* dataset seperti yang di gambarkan pada Histogram berikut ini.



Gambar 4. Scatter Matrix

3.2 Pengolahan Data Awal

Penulis melakukan pengolahan data awal agar mendapatkan data yang berkualitas untuk melakukan penelitian ini yaitu ada beberapa teknik yang dapat dilakukan yaitu sebagai berikut:

a. Data Validasi, fungsinya melaukan identifikasi dan menghapus data yang ganjil atau outer/noise, data tersebut yang merupakan data-data tidak konsisten.
b. Data Integrasi dan Transformasi, fungsinya untuk meningkatkan hasil penelitian agar lebih efisiensi dengan menggunakan algoritma yang telah di pilih dari peneitian.

c. Data Reduksi dan Dikrisasi fungsinya untuk mendapatkan dataset yang digunakan agar sesuai dengan atribut dan record yang cocok dan pas sehingga dapat di sebut degan data yang informatif.

3.3 Metode

Pada penelitian ini sebelumnya menggunakan beberapa percobaan dengan Algoritma Classification diantaranya LinearDiscriminantAnalysis, KNN dan CART. Dari beberapa algoritma tersebut akan dibandingkan dan diperoleh salah satu metode dengan hasil terbaik.

Pengujian model menggunakan Cross Validation, Evaluasi dengan Confusion Matrix dan kurva ROC sehingga dihasilkan akurasi dari metode tersebut. Lalu akan dilakukan komparasi terhadap metode trsebut sehingga didapatkan algoritma yang akurat untuk memprediksi *Bike Share Daily*.

3.1.1 Cara Kerja Metode LDA

Metode Diskriminan dapat menghasilkan interpretasi hasil analisis yang mudah dipahami selain itu analisis dari metode ini juga sensitive terhadap kehadiran pencilan [9]. Sehingga hasil dari tiap kelas selalu memiliki selihsih rata-rata besar sehingga terdapat beberapa variasi kecil untuk masing-masing kelas dengan kombinasi linier yang dapat dipisahkan untuk target kelas klasifikasi, perhitungan rumusnya sebagai berikut:

$\boldsymbol{Y\sim = W'X}$ ……………………………………………... (2)

Yang dimaksud dengan rumus tersebut yaitu merupakan penentu sebagai kombinasi untuk melakukan klasifikasi satu kolom dari matriks W, lalu selanjutnya:

$\boldsymbol{Yj = (W^{*})'X = W_{1^{*}}X_{1j} + w_{2^{*}} + X_{2j} + \cdots + W_{d}^{*}}$ …… (2) , untuk fungsi skornya berikut:

$\boldsymbol{\alpha = \frac{W'S_BW}{W'S_wW}}$, ………………………………………... (3)

dengan,

$$\boldsymbol{S_B = \sum_{i=1}^{k} d\,(X_i - X)(X_i - X)'}$$

$$\boldsymbol{S_w = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_i - X)(X_i - X)'}$$

3.1.2 Cara Kerja Metode k-NN

Metode k-NN merupakan kelompok instance-based-learning. Metode k-NN juga merupakan salah satu teknik yang dinamakan lazy-learning. K-NN dapat dilakukan dengan pencarian nilai dari kelompok k objek pada data training yang mirip atau paling depat dengan objek yang baru atau data testing [16].

Metode k-NN dapat menggunakan rumus sebagai berikut:

$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ ………………………….. (4)

Metode k-NN menentukan parameter nilai K dengan jumalah tetangga yan paling dekat dengan Parameter K pada nilai testing kemudian di tentukan berdasarkan nilai K yang optimum sehingga diperoleh dengan mencoba termenerus. Untuk menghitung nilai kuadrat jarak Euclid dari masing-masing objek terhadap data dari sampel yang telah di berikan. Mengurutkan beberapa objek ke kelompok yang sudah mempunyai jarak Euclid terkecil. Dengan mengumpulkan kategori dari nilai Y yang termasuk kategori mayoritas, maka dapat diprediksi oleh nilai query instance yang sudah di hitung sebelumnya.

3.1.3 Cara Kerja Metode CART

Klasifikasi CART terdiri dari 3 tahap yaitu : pemilihan pemilah, yang dimaksudkan disitu adalah ketergantungan nilai yang asalnya dari suatu variable yang independent [17]. Variabel yang termasuk independen kontinu *Xj* maka masuk dalam ruang sampel ukuran *n* dan terdapat *n* yaitu nilai amatan sampel yang beda. Sedangkan untuk nilai *Xj* merupakan variable dimana kategori nominalnya bertaraf L, kemudian diperoleh pemilihan sebanyak $\boldsymbol{2^{L-1} - 1}$. Namun jika variable *Xj* merupakan kategori ordinal maka harus dilakukan pemilihan yang memungkinkan.

Metode Pemilihan yang sangat sering digunakan yaitu indeks Gini, fungsi dari indeks Gini yaitu :

$\boldsymbol{i(t) = \sum_{i\neq j} p(i|t)p(j|t)}$…………………..…………(5)

Yang dimaksud dengan rumus $\boldsymbol{i(t)}$ yaitu fungsi keheterogenan indeks gini, *p(i|t)* dengan proporsi kelas *i* pada simpul *t*, dan p(j|t) adalah proporsi kelas *j* pada simpul *t*. *Goodness of split* yaitu suatu evaluasi pemilahan pemilah *s* pada simpul *t*. *Goodness of split* ϕ *(s,t)* didefinisikan sebagai penurunan keheterogenan.

$\boldsymbol{\emptyset(s,t) = \Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R)}$……..(6)

Pengembangan pohon dengan melakukan pencarian ke semua yang kemungkinan pemilah pada simpul $\boldsymbol{t_1}$ sehingga ditemukan pemilah s* yang memberikan nilai penurunan keheterogenan tertinggi yaitu:

$\boldsymbol{\Delta(i(s^{*},t_1) = max_{s\in S}\Delta i(s,t_1)}$……………………………..…………*(7)*

dengan $\boldsymbol{\emptyset(s,t)}$ merupakan kriteria *goodness of split*, $\boldsymbol{P_L i(t_L)}$ yaitu proporsi yang dapat mengapati dari semua simpul *t* lalu menuju simpul kiri, dan $\boldsymbol{P_R i(t_R)}$ yaitu proporsi dengan pengamatan yang dimulai dari simpul *t* lalu menuju simpul kanan. Tahap kedua yaitu penentuan simpul terminal, simpul *t* dapat dijadikan simpul terminal jika simpul tidak terdapat penurunan keheterogenan yang artinya pada pemilahan, hanya mempunyai nilai satu pengamatan *(n=1)* pada setiap simpul anak atau adanya pembatasan minimum *n* dan adanya batasan jumlah level. Tahap ketiga yaitu penandaan label tiap simpul terminal berdasar aturan jumlah anggota kelas terbanyak, yaitu:

$\boldsymbol{p(j|t) = max_j\, p(j|t) = max_j \frac{N_J(t)}{N(t)}}$……………………………………………..(8)

dengan $\boldsymbol{p(j|t)}$ yaitu jumlah proporsi kelas *j* pada simpul *t*, $\boldsymbol{Nj(t)}$ yaitu jumlah yang dilakukan pengamatan fi kelas *j* pada simpul *t*, dan $\boldsymbol{N(t)}$ yaitu jumlah pengamatan yang ada pada simpul *t*. Label kelas simpul terminal *t* yaitu $\boldsymbol{j_0}$ yang selalu memberi nilai dengan dugaan kesalahan klasifikasi pada simpul t dengan nilai besar. Proses dari pembentukan pohon klasifikasi berhenti ketika ada satu pengamat dalam tiap simpul anak dalam nilai *n*, semua pengamatan yang dilakukan dalam setiap simpul anak identik, lalu adanya batasan dari umlah level/kedalaman pohon maksimal. Setelah semua terbentuk pohon sudah maksimal kemudian selanjutnya yaitu pemangkasan pohon dimana dapat mencegah pembentukan pohon klasifikasi yang berukuran sangat besar dan kompleks, lalu bisa memperoleh ukuran yang layak berdasarkan *cost complexity prunning*, kemudian rumus besarnya *resubtitution estimate* pohon *T* pada parameter kompleksitas α yaitu :

$$\boldsymbol{R_a(T) = R(T) + a|T|}\text{.....(9)}$$

Dengan $\boldsymbol{R_a(T)}$ yaitu resubtitution suatu pohon *T* pada kompleksitas $\boldsymbol{a, R(T)}$ yaitu resubstitution estimate, bagi $\boldsymbol{a}$ yang merupakan $\boldsymbol{Parameter\ cost - complexity}$ dari satu simpul akhir pada pohon T, dan $\boldsymbol{|T|}$ dengan dilakukan banyak dari simpul terminal pohon *T*.

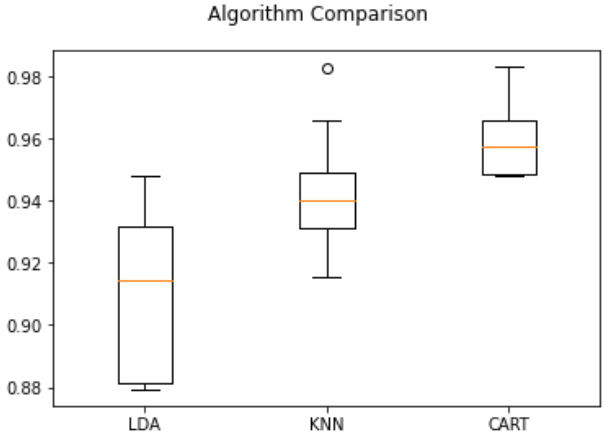
Cost complexity prunning menentukan pohon bagian $\boldsymbol{T(a)}$ yang dapat meminimumkan $\boldsymbol{R_a(T)}$ pada seluruh pohon bagian untuk setiap nilai *T*. Nilai parameter kompleksitas *T* akan secara perlahan meningkat selama proses pemangkasan. Selanjutnya pencarian pohon bagian $\boldsymbol{T(a) < T_{max}}$ yang dapat meminimumkan $\boldsymbol{R_a(T)}$ yaitu :

$$\boldsymbol{R_a((T_{(a)})) = min_{Tmax} R_a(T)}\text{.....(10)}$$

4. Hasil dan Pembahasan

Pada tahap ini peneliti melakukan eksperimen dan pengujian dengan beberapa metode yaitu Classification diantaranya Kneighbors Classifier, Linear Discriminant Analysis dan Classification And Regression Trees.

Pada penelitian ini menggunakan *Global Feature Exctraction* dengan beberapa diperoleh hasil akurasi seperti gambar berikut.



Gambar 6. Algorithm Comparison

Berdasarkan hasil diatas penulis juga melakukan percobaan dari dataset yanga da yaitu dengan melakukan perhitungan dengan algoritma LDA, KNN dan CART dengan hasil LDA menunjukkan 93% kemudian algoritma KNN menunjukkan hasil 95% dan algoritma CART menunjukkan 96% untuk akurasinya. Sehingga dapat disimpulkan bahwa Klasifikasi *Bike Share Daily Dataset* menggunakan berdasarkan gambar pada *Global Feature Extraction* untuk Akurasi tertinggi diperoleh pada CART dengan hasil akurasi sebesar 96%.

5. Kesimpulan

Dalam penelitian ini, penulis menggunakan media pendekatan pembelajaran mendalam dengan menerapkan beberapa metode yang telah di tampilkan dengan motode CART yaitu menunjukkan hasil akurasi 96%. Dengan menunjukkan arsitektur perkiraan distribusi di bawah berbagai sepeda dengan spatiotemporal variable, pendistribusian penggunaan dan waktu.

Makalah ini mengeksplorasi peramalan distribusi jangka pendek di bawah sistem berbagi sepeda tanpa dermaga melalui pendekatan spatiotemporal baru dalam arsitektur ujung ke ujung. Peramalan distribusi sepeda real-time yang akurat dapat memberikan saran bagi pengguna untuk merencanakan strategi perjalanan dengan lebih baik dan memenuhi kebutuhan perjalanan. Selain itu, ini juga berkontribusi pada penerapan strategi redistribusi dinamis untuk membuat sepeda lebih seimbang didistribusikan sesuai dengan tuntutan. Namun, model tidak dapat menjelaskan efek dari hubungan penawaran dan permintaan pada distribusi sepeda dan keseragaman melalui kerangka teoritis. Di masa depan, penulis berharap untuk membangun model teoretis tentang masalah ini.

Daftar Pustaka

[1] C. Jiang, Z. Wang, R. Wang, and Y. Ding, "Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending," *Ann. Oper. Res.*, vol. 266, no. 1–2, pp. 511–529, 2018, doi: 10.1007/s10479-017-2668-z.

[2] W. El-assi, M. S. Mahmoud, and P. K. N. Habib, "Effects of Built Environment and Weather on Bike Sharing Demand : A Station Level Analysis of Commercial Bike Sharing in Toronto."

[3] S. V E and Y. Cho, "A rule-based model for Seoul Bike sharing demand prediction using weather data," *Eur. J. Remote Sens.*, vol. 53, no. sup1, pp. 166–183, 2020, doi: 10.1080/22797254.2020.1725789.

[4] V. Albuquerque, F. Andrade, J. Ferreira, M. Dias, and F. Bacao, "Bike-sharing mobility patterns: a data-driven analysis for the city of Lisbon," *EAI Endorsed Trans. Smart Cities*, vol. 5, no. 16, p. 169580, 2018, doi: 10.4108/eai.4-5-2021.169580.

[5] R. K. Bock *et al.*, "Methods for multidimensional event classification: A case study using images from a Cherenkov gamma-ray telescope," *Nucl. Instruments Methods Phys. Res. Sect. A Accel.*

Spectrometers, Detect. Assoc. Equip., vol. 516, no. 2–3, pp. 511–528, 2004, doi: 10.1016/j.nima.2003.08.157.

[6] H. Amalia and E. Evicienna, "Komparasi Metode Data Mining Untuk Penentuan Proses Persalinan Ibu Melahirkan," *J. Sist. Inf.*, vol. 13, no. 2, p. 103, 2017, doi: 10.21609/jsi.v13i2.545.

[7] J. R. Saura, P. Palos-Sanchez, and A. Grilo, "Detecting indicators for startup business success: Sentiment analysis using text data mining," *Sustain.*, vol. 11, no. 3, pp. 1–14, 2019, doi: 10.3390/su11030917.

[8] V. Kotu and B. Deshpande, *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. Morgan Kaufmann, 2015.

[9] Y. N. N. Nanda Arista Rizki , Wasono, "Comparison of Classification Models Between Linear Discriminant Analysis and k-Nearest Neighbor for Students Majoring Data of Madrasah Aliyah Negeri in Samarinda," *Pros. Semin. Nas. Lingkung. Lahan Basah*, vol. 4, no. April, pp. 562–565, 2019.

[10] M. A. N. Febriansyach, F. Rashif, G. I. P. Nirvana, and N. A. Rakhmawati, "Implementasi LDA untuk Pengelompokan Topik Tweet Akun Bot Twitter bertagar #covid-19," *CogITo Smart J.*, vol. 7, no. 1, p. 170, 2021, doi: 10.31154/cogito.v7i1.299.170-181.

[11] Isman, Andani Ahmad, and Abdul Latief, "Perbandingan Metode KNN Dan LBPH Pada Klasifikasi Daun Herbal," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 3, pp. 557–564, 2021, doi: 10.29207/resti.v5i3.3006.

[12] I. A. Nikmatun and I. Waspada, "Implementasi Data Mining untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor," *J. SIMETRIS*, vol. 10, no. 2, pp. 421–432, 2019.

[13] M. M. Ghiasi, S. Zendehboudi, and A. A. Mohsenipour, "Decision tree-based diagnosis of coronary artery disease: CART model," *Comput. Methods Programs Biomed.*, vol. 192, p. 105400, 2020, doi: 10.1016/j.cmpb.2020.105400.

[14] A. Sururi, "Inovasi Kebijakan dalam Perspektif Administrasi Publik Menuju Terwujudnya Good Public Policy Governance," vol. 12, pp. 14–31, 2017, doi: 10.31227/osf.io/6djph.

[15] R. Rosliana and R. Loisa, "Strategi Cyber Public Relatons dalam Memanfaatkan Media Sosial untuk Membangun Citra Perusahaan," *Prologia*, vol. 2, no. 2, p. 480, 2019, doi: 10.24912/pr.v2i2.3733.

[16] F. D. Wahyudi, D. Remawati, and P. Harsadi, "Sistem Pakar Deteksi Kerusakan Mesin Bubut Dengan Metode Knn," *J. Teknol. Inf. dan Komun.*, vol. 6, no. 2, pp. 7–13, 2019, doi: 10.30646/tikomsin.v6i2.370.

[17] A. Dessy, "Perbandingan Ketepatan Klasifikasi Antara Metode Regresi logistik dan Klasifikasi Pohon," 2015.