 ISBN 978-623-5645-10-0

Klasifikasi User yang Berpotensi Melakukan Pembelian Barang Online Menggunakan Algoritme Weighted K-Nearest Neighbor

Valentina Yohana Senduk [a,*], Eko Hari Parmadi [b]

[a] Universitas Sanata Dharma, Yogyakarta
[b] Universitas Sanata Dharma, Yogyakarta
*hari@usd.ac.id

Abstract
Electronic commerce is an exchange using technology as an intermediary between various parties (individuals or organizations) as well as electronic-based intra or interorganizational exchange activities. The most commonly used interface is the web browser. Visitor's behavior on a web site can be analyzed using a classification algorithm on the data order of site visits or commonly known as web clickstream. In this study the weighted k-nearest neighbor algorithm is used to classify users who have the potential to make online based purchase on public data on a web clickstream data of an e-commerce. The weighted k-nearest neighbor method makes the classification based on the label that has the highest number of distance weights at the nearest k-neighbor from the classified data. The output of this system is the classification of users who have the potential to purchase goods online and those who do not. Researchers conducted an attribute selection experiment with variations in k and variations in the number of attributes based on the results of rank information gain and obtained the 2 most influential attributes in the classification namely Page Values and Exit Rates. Testing to find the most optimal k was done on 7,632 data using 3-fold cross validation and produced the highest accuracy of 86.2028% at k = 65.

Keywords : *e-commerce, web clickstream, weighted k-nearest neighbor, classification*

Abstrak
Perdagangan elektronik merupakan pertukaran yang menggunakan teknologi sebagai perantara antara berbagai pihak (individu atau organisasi) serta kegiatan pertukaran intra atau antar organisasi berbasis elektronik. Antar muka yang paling sering digunakan adalah web browser. Perilaku pengunjung di situs web dapat dianalisis menggunakan algoritma klasifikasi terhadap data urutan kunjungan halaman situs atau biasa dikenal sebagai web clickstream. Pada penelitian ini digunakan algoritma weighted k-nearest neighbor untuk mengklasifikasikan user yang berpotensi melakukan pembelian barang online berdasarkan data publik web clickstream sebuah e-commerce. Metode weighted k-nearest neighbor melakukan klasifikasi berdasarkan label yang memiliki jumlah bobot jarak terbesar pada k-tetangga terdekat dari data yang diklasifikasi. Keluaran dari sistem ini adalah klasifikasi user yang berpotensi melakukan pembelian barang online dan mana yang tidak. Peneliti melakukan percobaan seleksi atribut dengan variasi k dan variasi jumlah atribut berdasarkan hasil perhitungan information gain dan memperoleh 2 atribut yang paling berpengaruh dalam klasifikasi yakni Page Values dan Exit Rates. Pengujian untuk mencari k yang paling optimal dilakukan pada 7.632 data menggunakan 3-fold cross validation dan menghasilkan akurasi tertinggi yakni 86.2028% pada k = 65.
Kata Kunci : e-commerce, web clickstream, weighted k-nearest neighbor, klasifikasi

1. Pendahuluan

Semakin berkembangnya teknologi informasi, khususnya internet turut mengubah sistem perdagangan yang ada. Sistem perdagangan yang dulunya dilakukan melalui pertemuan langsung antara penjual dan konsumen atau pembeli, kini dapat dilakukan melalui dunia maya melalui perdagangan secara elektronik. Pengelola bisnis elektronik hanya memperoleh data dari hasil browsing pengunjung di situs webnya untuk mengklasifikasikan mana pengunjung yang akhirnya membeli barang dan mana yang tidak . Di sisi lain, melonjaknya penggunaan perdagangan elektronik berimbas pada besarnya data yang diperoleh. Himpunan data ini dapat diolah untuk menghasilkan pengetahuan untuk membantu pengelola bisnis elektronik merencanakan strategi pemasaran yang lebih baik. Salah satunya

dengan menggunakan algoritma klasifikasi untuk menganalisis data *web clickstream* pada situs web. Web *clickstream* sendiri merupakan urutan halaman web yang diminta oleh pengguna termasuk di antaranya *user session* yang menjelaskan halaman web yang dilihat oleh satu pengguna selama satu periode ketika masuk ke web. Perilaku pengunjung di situs web dapat diprediksi dengan menganalisis data yang ada pada urutan kunjungan halaman situs. Pada saat browsing, setiap kali pengguna menautkan ke situs web, server melacak semua tindakan pengguna *dalam file log. User session* ini bisa berisi halaman dari lebih dari satu situs. Selain itu ada pula server session, atau session, yang merupakan kumpulan halaman untuk satu situs tertentu selama user session. Halaman-halaman ini juga biasa disebut sebagai sebuah visit. Penelitian ini menggunakan Dataset yang digunakan dalam penelitian ini diperoleh dari situs *UCI Repository Machine Learning* yaitu *Online Shoppers Purchasing Intention Dataset* yang terdiri dari 12.330 data dengan 18 atribut sudah termasuk label, yakni TRUE yang menandakan pengunjung yang berakhir melakukan pembelian barang online dan FALSE yang menandakan pengunjung yang tidak melakukan pembelian barang online. Metode yang digunakan dalam penelitian ini adalah modifikasi dari k-nearest neighbor yakni weighted k-nearest neighbor. Masalahnya adalah menetukan atribut-atribut yang berpengaruh, nilai k serta besar akurasi yang dihasilkan algoritma weighted k-nearest neighbor dalam mengklasifikasikan pengunjung yang berpotensi melakukan pembelian barang online berdasarkan data *web clickstream.*

2. Kajian Literatur

Knowledge Discovery in Database (KDD)

Penambangan data *(data mining)* adalah proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual. (Kusrini &Luthfi, 2009). Istilah data mining dan *Knowledge Discovery in Database* (KDD) sering digunakan secara bergantian untuk menjelaskan proses penggalian informasi yang tersebunyi dalam suatu basis data yang besar.

Salah satu tahapan dalam keseluruhan proses KDD adalah data mining (Han, Jiawei, dkk. 2012). Proses KDD terdiri dari 7 tahap yaitu :

a. Pembersihan data *(data cleaning)*
   Tahap data cleaning dilakukan untuk membersihkan *noise* dan data yang inkosisten pada umumnya di tahap ini juga dilakukan pembersihan data missing values namun data set yang digunakan pada penelitian ini tidak mengandung *missing value*.

b. Integrasi data *(data integration)*
   Tahap ini akan dilalukan penggabungan data. Data dari bermacam-macam tempat penyimpanan data akan digabungkan ke dalam suatu tempat penyimpanan data yang koheren.

c. Seleksi data *(data selection)*
   Pemilihan (seleksi) adalah proses memilih data atau atribut yang relavan. Pada tahap ini dilakukan analisis korelasi atribut data. Atribut-atribut data tersebut dicek apakah relavan atau dilakukan penambangan data.

d. Transformasi data *(data transformation)*
   Transformasi adalah proses yang dilakukan untuk mengubah bentuk data menjadi bentuk yang sesuai untuk digunakan. Proses ini dilakkan untuk mengubah data di atribut yang belum numeric menjadi data numerik.

e. Penambangan data *(data mining)*
   Data mining adalah proses mencari pola atau infromasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu.

f. Evaluasi Pola *(pattern evaluation)*
   Dalam tahap ini hasil dari teknik data mining berupa pola yang khas maupun model prediksi dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai.

g. Presentasi pengetahuan *(knowledge presentation)*
   Pada langkah ini informasi yang sudah ditambang akan divisualisasikan dan direpresentasikan kepada pengguna. Langkah 1 sampai 4 merupakan langkah praproses data dimana data akan disiapkan terlebih dahulu selanjutnya dilakukan penambangan.

Klasifikasi pada Penambangan Data

Pada penambangan data terdapat berbagai jenis algoritma yang digunakan yakni salah satunya adalah algoritma untuk klasifikasi yang merupakan teknik untuk mengklasifikasi sebuah data baru untuk mengelompokkannya ke kelompok yang telah didefinisikan.

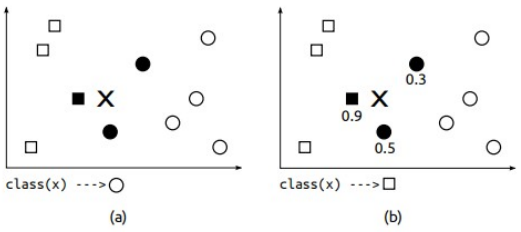
 ISBN 978-623-5645-10-0

Proses klasifikasi didasarkan pada empat komponen (Gorunescu, 2011) :

a. Kelas
   Variabel dependen yang berupa kategorikal yang mempresentasikan label yang terdapat pada objek. Contoh : resiko kredit, jenis gempa.
b. *Predictor*
   Variabel independen yang direpresentasikan oleh karakteristik data. Contoh: tabungan, aset, gaji.
c. *Training dataset*
   Set data yang berisi nilai dari kedua komponen di atas yang digunakan untuk menentukan kelas yang cocok berdasarkan *predictor.*
d. *Testing dataset*
   Set data yang akan diklasifikasikan oleh model yang telah dibuat. Akurasi klasifikasi nantinya akan dievaluasi.

Weighted k-Nearest Neighbor

Algoritma *k-Nearest Neighbor* menggunakan perhitungan kemiripan data baru (data *testing*) dengan data sebelumnya yang sudah memiliki label atau kelas (data *training*) sebagai nilai prediksi dari sampel uji yang baru. Perhitungan kemiripan dilakukan dengan menghitung jarak antar tetangga. Pada umumnya, jarak antar tetangga pada k-NN dihitung dengan *Eucledian distance.* Penentuan label data baru sendiri akan dilakukan melalui *voting* label mayoritas dari k-tetangga terdekat.



Gambar 1 Ilustrasi (a) k-NN dan (b) w-KNN
Sumber : Bichego & Loog (2016)

Masalah bisa terjadi ketika kelas terdekat mempunyai jarak yang bervariasi. Seperti pada gambar 1 (a) class(x) diklasifikasikan sebagai lingkaran karena dengan algoritma k-NN jumlah kelas terbanyak pada tetangga terdekat adalah lingkaran. Pada kenyataannya, class(x) yang sebenarnya adalah kotak. Cara untuk mengatasi ini adalah dengan pembobotan tetangga terdekat oleh jaraknya (Prasetyo, 2014). Metode pembobotan yang paling sering digunakan adalah dengan menginverskan jarak tiap k-tetangga terdekat. Metode k-Nearest Neighbor dengan pembobotan tetangga terdekat disebut *Weighted k-Nearest Neighbor* (W-kNN).

Berikut merupakan algoritma *weighted k-nearest neighbor* (Schliep, K. P, 2004) :

1. Menentukan parameter k (k adalah jumlah tetangga paling dekat yang akan disertakan dalam penentuan kelas).
2. Menghitung kuadrat jarak *Euclidean* setiap objek data latih terhadap data uji yang diberikan. Rumus untuk mendapatkan jarak antar data dengan jarak *Euclidean* dapat dihitung menggunakan persamaan (1) berikut ini :

$$d_i = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (1)$$

   dimana :
   d_i = jarak *Euclidean* variabel ke-i
   i = variabel data (i = 1,2,3,...,n)
   n = dimensi data
   x = data uji
   y = data latih

3. Mengurutkan hasil jarak *euclidean* dari terkecil ke terbesar.
4. Mengambil sejumlah k-tetangga terdekat.
5. Memberikan bobot pada masing-masing jarak pada k-tetangga terdekat menggunakan fungsi inversi. Pembobotan dengan fungsi inversi dapat dihitung menggunakan persamaan (2) berikut ini :

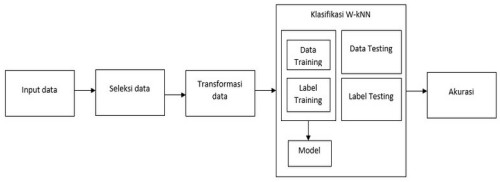
$$w_{(i)} = \frac{1}{¿ d i \vee ¿¿} \quad (2)$$

   dimana :
   $w_{(i)}$ = bobot variabel ke-i
   d_i = jarak *Euclidean* variabel ke-i
   i = variabel data (i = 1,2,3,...,n)
6. Menjumlahkan bobot tiap kelas pada k-tetangga terdekat.
7. Menggunakan kelas pada k-tetangga terdekat yang memiliki bobot terbesar sebagai label data uji.

3. Metode Penelitian

Pada sistem ini terdapat 5 proses yaitu input data, seleksi data, transformasi data, klasifikasi, dan penghitungan akurasi. Alur sistem secara umum dapat dilihat pada gambar 3.1 berikut :

 ISBN 978-623-5645-10-0



Gambar 2. Gambaran Umum Sistem

Data yang digunakan dalam penelitian ini adalah data publik yang diperoleh dari situs UCI *Repository Machine Learning* yaitu *Online Shoppers Purchasing Intention Dataset*. *Dataset* ini terdiri dari vektor fitur yang berasal dari 12.330 *session* yang dibentuk sehingga setiap sesi menjadi milik pengguna yang berbeda dalam periode 1 tahun guna menghindari kecondongan pada kampanye tertentu, periode tertentu, hari khusus, atau profil pengguna. Terdapat 18 atribut yaitu 10 atribut numerikal dan 7 atribut kategorikal dan satu kelas label. Label ada dua yakni FALSE yang menandakan pengunjung yang berakhir tidak membeli, dan TRUE yang menandakan pengunjung yang berakhir membeli. Penjelasan mengenai atribut yang digunakan dapat dilihat pada tabel 1

Tabel 1. Atribut Data Penelitian

No.	Atribut	Keterangan dan Nilai
1	Administrative	Jumlah kunjungan pengunjung ke halaman *Administrative*
2	Administrative Duration	Total waktu yang dihabiskan di halaman *Administrative*
3	Informational	Jumlah kunjungan pengunjung ke halaman *Informational*
4	Informational Duration	Total waktu yang dihabiskan di halaman *Informational*
5	Product Related	Jumlah kunjungan pengunjung ke halaman *Product Related*
6	Product Related Duration	Total waktu yang dihabiskan di halaman *Product Related*
7	Bounce Rates	Persentase pengunjung yang masuk ke situs dari halaman tersebut dan kemudian keluar *("bounce")* tanpa memulai permintaan lain ke *server analytics* selama *session* tersebut.
8	Exit Rates	Jumlah untuk semua tampilan halaman ke halaman, persentase yang terakhir dalam sesi
9	Page Values	Nilai rata-rata untuk halaman web yang dikunjungi pengguna sebelum menyelesaikan transaksi
10	Special Day	Jarak waktu pengunjung mengunjungi situs web ke hari khusus
11	Month	Bulan saat pengunjung mengunjungi situs web
12	Operating Systems	Sistem operasi yang digunakan pengunjung
13	Browser	Jenis *browser* yang digunakan pengunjung
14	Region	Wilayah sesuai sistem *browser* yang digunakan pengunjung
15	Traffic Type	*Traffic Type*
16	Visitor Type	Karakteristik pengguna, *New_visitor* merupakan pengunjung yang baru pertama kali mengunjungi situs web, *Returning_visitor* merupakan pengunjung yang sudah pernah mengunjungi situs web sebelumnya
16	Weekend	Nilai *boolean* yang mengindikasikan apakah hari pengunjung mengunjungi situs web adalah akhir pekan
18	Revenue	Label kelas yang bernilai TRUE apabila pengunjung akhirnya membeli barang di situs web, FALSE apabila pengunjung tidak membeli barang.

Seleksi Data

Pada tahap ini, akan dilakukan seleksi data untuk memilih atribut yang dibutuhkan dan menghapus atribut yang kurang relevan untuk penelitian. Kolom *Revenue* menjadi kelas yang digunakan untuk klasifikasi dan bernilai TRUE dan FALSE. Data yang digunakan pada tahap ini sejumlah 7.632 dengan rincian 1.908 bernilai 1 atau TRUE, dan 5.724 bernilai 0 atau FALSE. Pengurangan data dilakukan untuk menyeimbangkan data dengan perbandingan 1:3 data yang bernilai TRUE:FALSE. Atribut-atribut pada data dirangking dengan metode *Information Gain* pada fitur di aplikasi *Weka* versi 3.8.3. Hasil penghitungan *information gain* dapat dilihat pada tabel 2 berikut ini :

Tabel 2 *Information Gain* Atribut Data

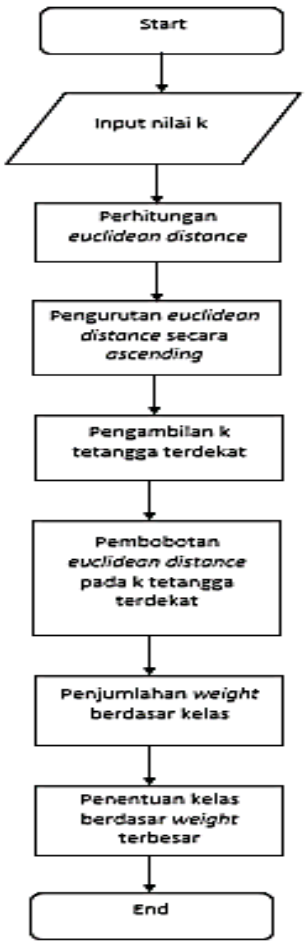
Rank	Nomor Kolom	Atribut
0.32221	9	Page Values
0.08439	8	Exit Rates
0.05715	6	Product Related

		Duration
0.0509	7	Bounce Rates
0.04728	5	Product Related
0.03232	1	Administrative
0.03214	11	Month
0.02855	2	Administrative Duration
0.02763	15	Traffic Type
0.01158	3	Informational
0.01129	4	Informational Duration
0.01085	16	Visitor Type
0.00861	10	Special Day
0.00558	12	Operating Systems
0.00143	17	Weekend
0.00112	13	Browser
0	14	Region

Berdasarkan hasil rangking atribut di atas, atribut Browser dan Region yang nilainya mendekati 0 tidak digunakan untuk klasifikasi. Tahap selanjutnya adalah transformasi data kemudian klasifikasi data dengan metode weighted k-nearest neighbor menggunakan atribut dengan rangking tertinggi kemudian ditambahkan satu per satu atribut dengan rangking tertinggi berikutnya untuk dilihat akurasinya. Pada tahap transformasi data dilakukan normalisasi pada data yang bertipe numerikal menggunakan metode normalisasi min-max dengan skala 0-1 dan transformasi data ke tipe numerikal untuk data bertipe kategorikal.

Klasifikasi dengan Weighted k-Nearest Neighbor

Flowchart klasifikasi dengan Weighted k-Nearest Neighbor dapat dilihat pada gambar 3. berikut ini (Schliep, K. P.,2004) :



Gambar 3. Klasifikasi dengan *Weighted k-Nearest Neighbor*

Fungsi yang digunakan untuk pembobotan hasil perhitungan jarak *euclidean distance* adalah fungsi inversi pada persamaan (2).

Tabel 3. Pemetaan langkah-langkah *W-kNN* dengan *function*

No.	Langkah	*Function*
1.	Input nilai k	wknn2.m
2.	Perhitungan *euclidean distance*	eucDistance.m
3.	Pengurutan *euclidean distance* secara *ascending*	wknn2.m
4.	Pengambilan k tetangga terdekat	wknn2.m
5.	Pembobotan *euclidean distance* pada k tetangga terdekat	wknn2.m
6.	Penjumlahan *weight* berdasar kelas	wknn2.m
7.	Penentuan kelas berdasar *weight* terbesar	wknn2.m

Confusion Matrix

Evaluasi hasil klasifikasi dilakukan dengan menghitung akurasi hasil klasifikasi berdasarkan confusion matrix. Representasi hasil klasifikasi pada confusion matrix menggunakan empat penentu untuk mencari akurasinya sebagaimana dapat dilihat pada tabel 3 berikut ini :

Tabel 4 *Confusion Matrix*

Label	***Positive***	***Negative***
Positive	*TP (True Positive)*	*FN (False Negative)*
Negative	*FP (False Positive)*	*TN (True Negative)*

Akurasi dihitung menggunakan persamaan berikut ini :

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} x\,100\,\% \quad (3)$$

dimana :

TP = jumlah positive yang diklasifikasikan sebagai positive

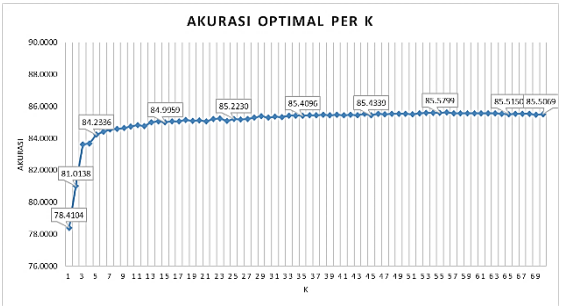
TN = jumlah negative yang diklasifikasikan sebagai negative

FP = jumlah negative yang diklasifikasikan sebagai positive

FN = jumlah positive yang diklasifikasikan sebagai negative

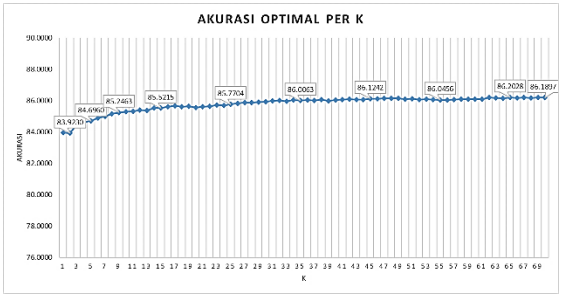
4. Hasil dan Pembahasan

Pengujian klasifikasi dengan 3-fold cross validation dilakukan pada data tanpa penyeimbangan sejumlah 12.330 data dan data setelah penyeimbangan sejumlah 7.632 data. Disini dicari nilai k dan jumlah atribut yang paling optimal untuk klasifikasi weighted k-nearest neighbor. Percobaan berhenti ketika hasil akurasi pada k mulai menurun.



Gambar 4. Nilai k dan akurasinya pada data tanpa penyeimbangan

Berdasarkan hasil percobaan pada data yang tanpa penyeimbangan, diperoleh akurasi tertinggi sebesar 85,5799 pada nilai k = 55.



Gambar 5. Nilai k dan akurasinya pada data dengan penyeimbangan

Hasil percobaan pada data dengan peyeimbangan didapat akurasi tertinggi sebesar 86,2028 pada nilai k = 65, atribut yang digunakan adalah *Page Values* dan *Exit Rates*.

5. Kesimpulan dan Saran

Kesimpulan

(1) Klasifikasi *user* yang berpotensi melakukan pembelian barang online terhadap 7.632 data web clickstream menggunakan metode *Weighted k-Nearest Neighbor* menghasilkan akurasi sebesar 86.2028% dengan model paling optimal pada k = 65

(2) Atribut yang berpengaruh dalam klasifikasi *user* adalah Page Values (Nilai rata-rata untuk halaman web yang dikunjungi pengguna sebelum menyelesaikan transaksi) dan *Exit Rates* (Jumlah untuk semua tampilan halaman ke halaman, persentase yang terakhir dalam sesi)

Saran

(1) Klasifikasi *user* dapat dikembangkan dengan metode lain

(2) Penyeimbangan data digunakan metode lain.

Daftar Pustaka

Bicego, M. & Loog, M.. (2016). *Weighted K-Nearest Neighbor revisited*. 1642-1647. 10.1109/ICPR.2016.7899872.

Gorunescu, Florin. (2011). *Data Mining : Concepts, Model, and Techniques.* Springer.

Han, Jiawei, dkk. (2012). *Data Mining : Concepts and Techniques*, Elsevier Inc.

Kusrini dan Luthfi, E.T.(2009). *Algoritma Data Mining*. Yogyakarta: ANDI.

Prasetyo, E. (2014). *Data Mining – Mengolah Data Menjadi Informasi Menggunakan Matlab.* Penerbit Andi : Yogyakarta.

Schliep, K. P. (2004). *Weighted k-nearest-neighbor techniques and ordinal classification*. Open Access LMU. https://doi.org/10.5282/ubm/epub.1769