PENERAPAN METODE K-NEAREST NEIGHBOR PADA PENENTUAN GRADE DEALER SEPEDA MOTOR

Henny Leidiyana
Program Studi Manajemen Informatika
Akademi Manajemen dan Informatika Bina Sarana Informatika
Jl. Dewi Sartika 77, DKI Jakarta
E-mail: henny.hnl@bsi.ac.id

Abstract— *The mutually beneficial cooperation is a very important thing for a leasing and dealer. Incentives for marketing is given in order to get consumers as much as possible. But sometimes the surveyor objectivity is lost due to the conspiracy on the field of marketing and surveyors. To overcome this, leasing a variety of ways one of them is doing ranking against the dealer. In this study the application of the k-Nearest Neighbor method and Euclidean distance measurement to determine the grade dealer. From the test results obtained by the value of accuracy of 64.03%.*

Keywords: *K-Nearest Neighbor, Cross Validation, Confusion matrix, ROC curve*

Intisari— Kerjasama yang saling menguntungkan adalah hal yang sangat penting bagi sebuah *leasing* dan *dealer*. Insentif bagi *marketing* diberikan agar mendapatkan konsumen sebanyak-banyaknya. Namun terkadang objektifitas *surveyor* hilang disebabkan oleh permainan di lapangan antara *marketing* dan surveyor. Untuk mengatasi hal tersebut, *leasing* melakukan berbagai cara salah satunya adalah melakukan peringkatan terhadap *dealer*. Dalam penelitian ini penerapan metode *k-Nearest Neighbor* dan pengukuran jarak *Euclidean* untuk menentukan grade *dealer*. Dari hasil pengujian diperoleh nilai keakuratan sebesar 64,03% .

Kata Kunci: *K-Nearest Neighbor, Cross Validation, Confusion matrix, ROC curve*

I. PENDAHULUAN

Dealer merupakan penyumbang konsumen bagi *leasing*. Dari hasil riset yang dilakukan penulis, setiap bulan lebih kurang 200 konsumen mengajukan kredit motor dimana konsumen mengajukan permohonannya melalui *dealer*. Oleh sebab itu kerjasama yang saling menguntungkan adalah hal yang sangat penting. Insentif bagi *marketing* diberikan agar mendapatkan konsumen sebanyak-banyaknya nya dan *surveyor* harus menjaga objektifitas dalam analisa kelayakan. Namun terkadang objektifitas hilang disebabkan oleh permainan di lapangan antara *marketing* dan surveyor. Berdasarkan informasi dari *leasing* dimana penulis melakukan riset, pada bulan Januari sampai Juni 2016 terjadi sebesar 92 konsumen mengalami permasalahan pada pembayaran kredit sepeda motornya.

Untuk mengatasi hal tersebut, *leasing* melakukan berbagai cara salah satunya adalah melakukan peringkatan terhadap *dealer*. Peringkat dealer setiap bulan ditinjau ulang agar selalu *update*, karena kondisi dealer tidak selalu stabil kualitasnya salah satunya disebabkan oleh permasalahan kenakalan oknum *surveyor* maupun *marketing* tadi. Peringkatan dilakukan oleh analis dengan memberikan grade bagi tiap *dealer* yang bekerjasama dengan *leasing*. Peringkatan dilakukan analis berdasarkan data dan menggunakan aplikasi Ms-Office. Peringkatan ini dimaksudkan agar *leasing* bisa menentukan komposisi *surveyor* yang ditempatkan di *dealer* dan besar target konsumen sekaligus mengendalikan kualitas konsumen.

Dalam penelitian ini penulis akan menerapkan metode *k-Nearest Neighbor* (K-NN) dalam penentuan grade dealer dengan melakukan pendekatan *data mining*.

II. BAHAN DAN METODE

Untuk penggalian data (*data mining*) dengan metode K-NN dalam menentukan grade *dealer*, data *training* didapat dari sebuah *leasing* kendaraan sepeda motor yang terletak di daerah Karawang, Jawa Barat. Data yang diperoleh adalah data seluruh dealer yang terletak di daerah Karawang yang bekerjasama dengan *leasing* dimana penulis melakukan riset.

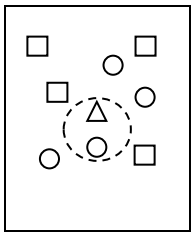
Dalam proses penggalian data, dilakukan beberapa tahapan, yang pertama yaitu membersihkan data, penerapan metode K-NN, dan

pengujian untuk mengukur unjuk kerja dari metode K-NN yang diterapkan, menggunakan metode pengujian *Cross Validation*, dan *Confusion Matrix*.

Data Mining (Witten, 2011) didefinisikan sebagai proses penemuan pola dalam data. dalam hal ini penulis melakukan tugas klasifikasi data mining. Dalam klasifikasi, data set tersedia lalu diamati, dan kelas target diketahui (Vercellis, 2011). Berbagai metode dalam klasifikasi dpt diterapkan. Pada penulisan ini digunakan metode K-NN.

Metode K-NN bekerja dengan cara mengukur kedekatan antara objek baru dengan objek lama untuk menentukan termasuk kelas mana objek baru tersebut (Gorunescu, 2011).

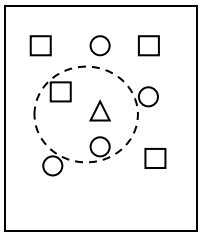
K-NN memiliki kelebihan, yang pertama yaitu relatif tidak sensitif terhadap error dalam dataset, dan yang kedua adalah K-NN dapat digunakan untuk memproses dataset berukuran besar (Myatt, 2007).



Sumber: Gorunescu, 2011

Gambar 1. Ilustrasi 1-nearest neighbor

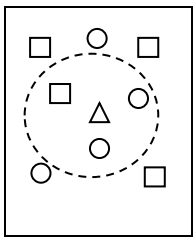
Pada ilustrasi Gambar 1 misalkan bentuk segitiga adalah objek yang akan diprediksi kelasnya. Jika nilai *K* sama dengan satu maka berarti hanya satu tetangga terdekat yang akan diperhitungkan jaraknya.



Sumber: Gorunescu, 2011

Gambar 2. Ilustrasi 2- nearest neighbor

Pada ilustrasi Gambar 2 jika nilai *K* sama dengan dua maka berarti dua tetangga terdekat yang akan diperhitungkan jaraknya. Sedangkan pada ilustrasi Gambar 3 jika nilai *K* sama dengan tiga maka berarti dua tetangga terdekat yang akan diperhitungkan jaraknya.



Sumber: Gorunescu, 2011

Gambar 3. Ilustrasi 3- nearest neighbor

Pada K-NN tentukan dulu jumlah objek terdekat yang diinginkan, disebut *K* objek. Misalkan *K*=5 lalu hitung jarak objek baru dengan dengan kelima (*K*=5) objek yang menjadi parameter tadi. Diantara lima objek tersebut mana yang paling dekat jaraknya maka objek baru akan masuk kelas yang sama dengan objek yang terdekatnya (Giudici,2009).

Pada metode K-NN, menentukan nilai K dan metode pengukuran jarak merupakan hal yang krusial. Ada banyak cara untuk mengukur jarak kedekatan antara objek baru dengan objek lama (data *training*), diantaranya *Euclidean distance* dan *Jaccard distance* (Myatt, 2007). Pada penulisan ini untuk mengukur jarak kedekatan objek lama dan objek baru menggunakan digunakan metode pengukuran jarak *Euclidean* (Myatt, 2011*)*. Rumus *Euclidean* untuk pengukuran jarak dua objek adalah:

$$d = \sqrt{x^2+y^2} \qquad \text{..... (1)}$$

Jika variable lebih dari dua maka rumus *Euclidean* untuk pengukuran jarak dua objek adalah:

$$d = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \qquad \text{..... (2)}$$

Untuk mengukur kinerja klasifikasi dalam penulisan ini digunakan *Cross validation,* yaitu pengujian yang dilakukan untuk memprediksi tingkat kesalahan (*error rate)*. Data *training* dibagi secara random ke dalam beberapa bagian dengan perbandingan yang sama kemudian *error rate* dihitung tiap bagian, selanjutnya hitung rata-rata seluruh *error rate* untuk memperoleh *error rate* secara keseluruhan (Witten,2011).

Metode *Confusion matrix* menggunakan tabel matriks. Jika data set hanya terdiri dari dua kelas, kelas yang satu dianggap sebagai positif dan yang lainnya negatif (Gorunescu,2011). Tabel 1 menunjukkan model *Confusion Matrix* dua Kelas.

Tabel 1 Model *Confusion Matrix* Dua Kelas

		Kelas Diprediksi	
		Ya	Tidak
Kelas Aktual	Ya	0	1
	Tidak	1	0

Sumber: Witten, 2011

Dalam suatu kasus misalkan bias saja kelas bernilai lebih dari dua, seperti dalam penelitian ini, atribut kelas grade memiliki 5 nilai, Yaitu A, B, C, D, E. Tabel 2 merupakan model *Confusion Matrix* untuk tiga Kelas.

Tabel 2 Model *Confusion Matrix* Dua Kelas

		Kelas Diprediksi		
		A	B	C
Kelas Aktual	A	0	1	1
	B	1	0	1
	C	1	1	0

Sumber: Witten, 2011

Setelah proses klasifikasi dengan metode K-NN dilakukan, berikutnya adalah mengukur kinerja dari klasifikasi yang dihasilkan dengan menghitung *sensitivity, specificity, pecision,* dimana rumusnya adalah:

$$Sensitivity = \frac{TP}{P} \quad \ldots\ldots (3)$$

$$Spesificity = \frac{TN}{N} \quad \ldots\ldots (4)$$

$$Precision = \frac{TP}{(TP+FP)} \quad \ldots\ldots (5)$$

$$Accuracy = \text{Sensitivity}\ \frac{P}{(P+N)} + \text{Sensitivity}\ \frac{N}{(P+N)} \quad \ldots\ldots (6)$$

Keterangan:
TP = jumlah *true positives*
TN = jumlah *true negatives*
P = jumlah objek positif
N = jumlah objek negatif
FP = jumlah *false positives*

True positives yaitu jumlah objek positif yang benar diklasifikasikan, *false positives* adalah jumlah objek negatif yang benar diklasifikasikan, *false negatives* adalah jumlah objek positif yang salah diklasifikasikan, *true negatives* adalah jumlah *objek* negatif yang salah diklasifikasikan.
Kemudian masukkan data uji ke dalam *confusion matrix*, setelah itu hitung nilai-nilai yang telah dimasukkan tersebut untuk menentukan *sensitivity* (*recall*), *specificity*, *precision* dan *accuracy*. *Sensitivity* digunakan untuk membandingkan jumlah TP terhadap jumlah objek yang positif sedangkan *specificity* yaitu perbandingan jumlah TN terhadap jumlah objek yang negatif

III. HASIL DAN PEMBAHASAN

Data yang digunakan sebanyak 81. Data *training* terdiri dari 6 atribut, dimana 5 atribut merupakan prediktor dan 1 atribut label. Atribut dealer, FID DP, FID tipe pada data *training* bernilai kategori dan atribut FID DP dan kontribusi dealer bernilai numerik, dan satu atribut kelas yaitu grade, seperti terlihat pada Tabel 3. FID (*First Installment Default*) yaitu angsuran pertama yang bermasalah. FID ini terjadi pada objek dengan DP, merk, tipe tertentu. Untuk mendapatkan data yang berkualitas, dilakukan *preprocessing*.

Tabel 3. Atribut dan Nilainya

NO	ATRIBUT	JENIS ATRIBUT DAN NILAINYA
1	Dealer	Kategori (nama dealer)
2	FID DP	Numerik
3	FID Merk	Kategori (Honda, Yamaha, Suzuki, Kawasaki)
4	FID Tipe	Kategori (Tipe-tipe tiap merk)
5	Kontribusi	Numerik
6	Grade	Label (A, B, C, D, E)

Sumber: *Leasing*, 2016

Pada Tabel 4 terlihat ada enam sampel *data training* dari 81 data yang diperoleh dari *leasing*. Data *training* merupakan data yang akan digunakan untuk proses klasifikasi dengan metode K-NN.

Tabel 4. Sampel data *training*

DEALER	FID DP	FID MERK	FID TIPE	KNTRB	GRADE
DEALER AM	500000	HONDA	BEAT POP	10	B
DEALER RD	500000	HONDA	REVO FIT	5	A
DEALER PL	500000	YAMAHA	MIO M3	10	B
DEALER WJ	500000	YAMAHA	MIOM3	5	A
DEALER MT	500000	SUZUKI	SATRIA	4	A
DEALER DY	500000	HONDA	BEAT POP	8	B

Sumber: *Leasing*, 2016

Untuk pengujian dalam penulisan ini diberikan contoh data *testing* seperti pada Tabel 5.

Pada Tabel 5 ada dua objek baru yang akan diprediksi kelasnya.

Tabel 5. Sampel data *testing*

DEALER	FID DP	FID MERK	FID TIPE	KNTRB	GRADE
DEALER A	500000	HONDA	BEAT POP	7	?
DEALER B	1000000	YAMAHA	MIO M3	10	?

Sumber: *Leasing*, 2016

Langkah pertama adalah melakukan pembersihan data dari *noise*, reduksi fitur/atribut (*feature reduction*) dan duplikasi. Dalam data traning Tabel 4, atribut dealer tidak diikutsertakan dalam perhitungan jarak karena tidak berpengaruh dalam penentuan klasifikasi grade.

Karena menghitung jarak menggunakan rumus *Euclidean* maka langkah berikutnya adalah mengubah data kategori ke dalam numeric. Dapat dilakukan dengan mengganti data dengan angka tertentu asalkan konsisten (Witten, 2011). Misalkan untuk atribut FID Merk, nilai-nilainya diubah sebagai berikut:

HONDA = 7
YAMAHA = 8
SUZUKI = 9

BEAT POP = 1
REVO FIT = 2
MIO M3 = 3
SATRIA = 4

Tabel 6 adalah hasil konversi data *training* dari Tabel 4 untuk atribut bernilai kategori.

Tabel 6. Konversi Data

DEALER	FID DP	FID MERK	FID TIPE	KNTRB	GRADE
1	500000	7	1	10	B
2	500000	7	2	5	A
3	500000	8	3	10	B
4	500000	8	3	5	A
5	500000	9	4	4	A
6	500000	7	1	8	B

Sumber: hasil olahan

Untuk data *testing* juga dilakukan konversi dengan nilai yang sama seperti pada Tabel 7.

Tabel 7. Sampel data *testing*

DEALER	FID DP	FID MERK	FID TIPE	KNTRB	GRADE
A	500000	7	1	12	?
B	1000000	8	3	5	?

Sumber: hasil olahan

Langkah kedua adalah melakukan perhitungan jarak. Pada penelitian ini nilai *K* yang digunakan adalah 5. Karena variable yang digunakan sebagai predictor ada empat, yaitu FID DP, FID merk, FID tipe, dan kontribusi maka rumus pengukuran jarak yang digunakan adalah:

$$d = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

Misalkan untuk objek pertama pada data *training* dihitung jaraknya dengan data *testing* sebagai berikut:

$$d = \sqrt{(500000-500000)^2+(7-7)^2+(1-1)^2+(10-12)^2}$$

$$= \sqrt{0+0+0+4}$$

$$= 2$$

Perhitungan seperti di atas dilakukan terhadap seluruh data dari data *training*. Hasilnya seperti pada Table 8. Terlihat bahwa objek Dealer 2 memiliki jarak paling besar dengan data objek baru Dealer A.

Tabel 8. Hasil Perhitungan Jarak

DEALER	JARAK	GRADE
1	2	B
2	6	A
3	1	B
4	4	A
5	3	A
6	4	B

Sumber: hasil olahan

Kemudian urutkan hasil berdasarkan jarak seperti pada Tabel 9 agar lebih terlihat kelompok berdasarkan parameter *K* nya. Kolom tiga pada Tabel 9 untuk memperjelas dari keenam objek data pada sample data *training* tersebut yang mana yang termasuk ke dalam lima tetangga terdekat. Dari enam data, lima data dengan jarak terdekat diperhitungkan. Sedangkan data dengan jarak terjauh tidak masuk ke dalam kelompok tetangga terdekat.

Untuk objek baru Dealer B dilakukan langkah yang sama dengan objek baru Dealer A.

Tabel 9. Hasil Perhitungan Jarak Terurut

DEALER	JARAK	Termasuk ke dalam 5-nearest neighbor	GRADE
3	1	Ya	B
1	2	Ya	B
5	3	Ya	A
4	4	Ya	A
6	4	Ya	B
2	6	tidak	A

Sumber: hasil olahan

Karena nilai *K* = 5 maka berarti dari 6 sample data *training* maka dealer 2 tidak termasuk dalam tetangga dengan jarak terdekat (*nearest neighbor*).

Setelah proses klasifikasi dengan metode K-NN dilakukan, berikutnya adalah mengukur kinerja dari klasifikasi yang dihasilkan dengan menghitung *sensitivity, specificity, pecision,* dan *accuracy* yang hasilnya tersaji pada table 10.

Tabel 10. *Confusion marix* hasil pengujian

accuracy: 64.03% +/- 10.71% (mikro: 64.20%)

	true A	true B	true C	true D	true E	class precision
pred. A	47	14	6	2	2	66.20%
pred. B	3	5	0	1	1	50.00%
pred. C	0	0	0	0	0	0.00%
pred. D	0	0	0	0	0	0.00%
pred. E	0	0	0	0	0	0.00%
class recall	94.00%	26.32%	0.00%	0.00%	0.00%	

Sumber: hasil pengujian

Pada table 10, terlihat bahwa klasifikasi dengan metode K-NN tingkat akurasinya adalah 64,03%.

IV. KESIMPULAN

Penentuan grade *dealer* oleh *leasing* dimana penulis melakukan riset saat ini dilakukan oleh analis menggunakan data dan aplikasi Ms-Office. Dalam penelitian ini dilakukan dengan pendekatan *data mining* yang terdiri dari tahapan *preprocessing*, pembuatan model dengan menerapkan metode K-NN dan pengukuran jarak *Euclidean* untuk menentukan grade *dealer*, dan pengujian. Metode K-NN bekerja dengan cara mengukur kedekatan antara objek baru dengan objek lama untuk menentukan termasuk kelas mana objek baru tersebut. Dari hasil pengujian diperoleh nilai keakuratan sebesar 64,03%

V. REFERENSI

Giudici, Paolo. 2009. Applied Data Mining for Business and Industry, Second Edition. UK: John Wiley & Sons

Gorunescu, Florin. 2011. *Data Mining: Concepts, Models, and Techniques*. Verlag Berlin Heidelberg : Springer

Myatt, Glenn J. 2007. Making Sense of Data, A Practical Guide to Exploratory Data Analysis and Data Mining. Hoboken, New jersey: : John Willey & Sons, Ltd.

Vercellis, Carlo. 2009. *Business Intelligent: Data Mining and Optimization for Decision Making*. Southern Gate, Chichester, West Sussex : John Willey & Sons, Ltd.

Witten, I. H., Frank, E., & Hall, M. A. 2011. *Data Mining: Practical Machine Learning and Tools*. Burlington : Morgan Kaufmann Publisher

BIODATA PENULIS

Henny Leidiyana, S.Kom, M.Kom. Jakarta, 12 Nopember 1975. Tahun 1998 lulus dari Program Strata Satu (S1) Jurusan Teknik Informatika Universitas Persada Indonesia YAI. Tahun 2011 lulus dari Program Strata Dua (S2) Jurusan Magister Ilmu Komputer STMIK Nusa Mandiri Jakarta. Jabatan Fungsional Akademik Asisten Ahli di AMIK BSI Jakarta. Telah menulis beberapa paper di Jurnal PIKSEL Universitas Islam 45 Bekasi, Jurnal-O Stmik Antar Bangsa, dan ICT Journal Bina Insani.