Pemilihan Pelanggan Potensial Dengan Melakukan Pemetaan Area Dengan Metode Algoritma K-NN dan K-Means Di Yamaha Nusantara Motor Purwokerto

Diwahana Mutiara Candrasari Hermanto 1*, Akhiles Frista Sugianto 2, Oskar Ika Adi Nugroho 3
Program Studi Teknik Informatika , STIKOM Yos Sudarso[1,2,3]
*Email : candrasari5860@stikomyos.ac.id [1], akhilesf@gmail.com [2]

Diterima:- . Disetujui:- . Dipublikasikan: Desember 2021

ABSTRAK

Penelitian ini mengusulkan sebuah clustering terhadap pelanggan potensial di Yamaha Nusantra Motor Purwokerto berdasarkan karakteristik konsumen. Clustering merupakan salah satu proses dari data mining yang bertujuan untuk mempartisi yang ada kedalam satu atau lebih cluster objek berdasarkan karakteristik yang milikinya. Metode yang digunakan dalam penelitian adalah K-Nearest Neighbor sebagai penentuan kelayakan data sedangkan K-Means digunakan sebagai clustering pelanggan. Nilai Davies Bouldin Index diteliti menggunakan rapidminer sedangakan purity validasi menggunakan Microsoft excel. Hasil menunjukan bahwa K-Means data training Davies Bouldin Index sebesar 0,266 dan Purity validasi sebesar 1,3351. Dalam K-Means data testing hasil menunjukan nilai Davies Bouldin Index sebesar 0,298 dan Purity validasi sebesar 0,6631.

Kata Kunci: K-Nearest Neighbor, K-Means, Davies Bouldin Index, Purity

ABSTRACT

This study proposes a clustering of potential customers at Yamaha Nusantra Motor Purwokerto based on consumer characteristics. Clustering is one of the processes of data mining that aims to partition existing objects into one or more clusters of objects based on their characteristics. The method used in this study is K-Nearest Neighbor as a determination of the feasibility of the data while K-Means is used as customer clustering. The value of the Davies Bouldin Index was examined using rapidminer while the purity validation was using Microsoft excel. The results show that the K-Means training data of the Davies Bouldin Index is 0.266 and the Purity validation is 1.3351. In the K-Means data testing the results show the Davies Bouldin Index value of 0.298 and the Purity validation of 0.6631.

Keywords: K-Nearest Neighbor, K-Means, Davies Bouldin Index, Purity

PENDAHULUAN

Kemajuan teknologi informasi sudah semakin berkembang pesat disegala bidang kehidupan. Banyak sekali data yang dihasilkan oleh teknologi informasi yang canggih, mulai dari bidang industri, ekonomi, ilmu dan teknologi serta berbagai bidang kehidupan lainnya. Pada proses penetuan pelanggan potensial pada umumnya menggunakan beberapa faktor.

Perusahaan Yamaha Nusantara Motor merupakan perusahaan yang bergerak dalam bidang otomotif seperti halnya penjualan kendaraan roda dua atau motor maupun service atau perbaikan motor. Perusahan Yamaha Nusantara Motor menjual berbagai jenis sepeda motor seperti Sport, Standard/Bebek dan Skuter Matic. Permasalahan yang muncul adalah dimana daerah yang paling potensial di Purwokerto

yang sangat produktif dalam penjualan. Sering kali dalam penentuan pemasaran masih kurang efektif karena daerah yang dilakukan saat pameran masih kurang berpotensi dalam segi penjualan. Dalam penelitian ini pengelompokan pelanggan potensial menggunakan 6 indikator yang digunakan dalam penentuan.

Penelitian ini menggunakan metode K-Nearest Neighbor dan K-Means. Dimana penggunaan data dibagi menjadi dua data yaitu data training dan data testing. Data training menggunakan 100 data sedangkan data testing menggunakan 400 data yang diproses menggunakan metode K-Nearest Neighbor dan K-Means. Setelah melakukan pehitungan tersebut akan dilakukan uji validasi menggunakan Davies Bouldin Index dan Purity Validasi.

METODE PENELITIAN

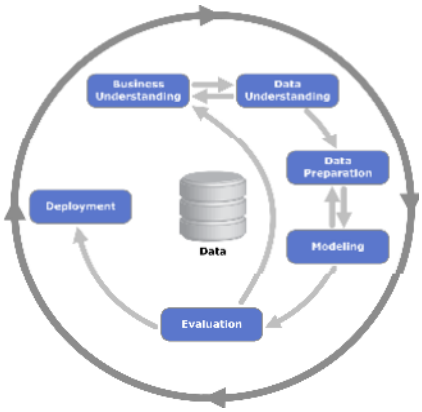
1. Data Mining

Data mining bukanlah suatu bidang yang sama sekali baru. Salah satu kesulitan untuk mendefinisikan data mining adalah kenyataan bahwa data mining mewarisi banyak aspek dan teknik dari bidang-bidang ilmu yang sudah mapan terlebih dahulu. Berawal dari beberapa disiplin ilmu. Data mining bertujuan untuk memperbaiki teknik tradisional sehingga bisa menangani(Fajrin & Maulana, 2018):

a. Jumlah data yang sangat besar
b. Dimensi data yang tinggi
c. Data yang heterogen dan berbeda sifat.

2. CRISP-DM

CRISP-DM merupakan metode yang menggunakan model proses pengembangan data yang banyak digunakan para ahli untuk memecahkan masalah. Proses penelitian ini mengacu pada enam tahap CRISP-DM yaitu pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi dan penyebaran(Setiawan, 2016). Proses tersebut dapat dilihat pada gambar 1.



Gambar 1 Framework CRISP-DM

3. Classification

Classification merupakan tindakan untuk memberikan kelompok pada setiap keadaan berisi sekelompok atribut, salah satunya adalah class attribute. Metode ini untuk menemukan model (atau fungsi) yang menggambarkan dan membedakan kelas data atau konsep yang bertujuan agar bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui(Annur, 2018).

4. K-Nearest Neighbor

Agoritma K-Nearest Neighbor (K-NN) adalah suatu metode yang menggunakan algoritma supervised. K-Nearest Neighbor (K-NN) merupakan algoritma untuk menentukan kelas objek data uji berdasarkan K objek pada data latih yang terdekat (mirip). Algoritma ini termasuk instance-based learning dan merupakan salah satu teknik lazy learning(Putra et al., 2017). Langkah-langkah proses menggunakan K- Nearest Neighbor adalah sebagai berikut(Nuari, Apriliyani, Kusrini, & Juwari, 2018):

a. Menghitung jarak Euclidean dengan rumus:

$$d(xi, xy) = \sqrt{\sum_{r}^{n} (a_r(x_i) - a_r(x_y))^2}$$

Keterangan :

d(xi, xy)	= Jarak *Euclidean*
x_i	= Record ke-i
x_y	= Record ke-y
a_r	= Data ke-r

b. Mengurutkan data jarak Euclidean
c. Menentukan jarak klasifikasi terdekat
d. Melakukan pengelompokan data dengan kesusuaian klasifikasi data

5. Clustering

Clustering termasuk ke dalam descriptive methods, dan juga termasuk unsupervised learning dimana tidak ada pendefinisian kelas objek sebelumnya. Sehingga clustering dapat digunakan untuk menentukan label kelas bagi data-data yang belum diketahui kelasnya. Konsep dasar dari clustering adalah mengelompokkan sejumlah objek ke dalam cluster dimana cluster yang baik adalah cluster yang memiliki tingkat kesamaan yang tinggi antar objek di dalam suatu cluster dan tingkat ketidaksamaan yang tinggi dengan objek cluster yang lainnya. Terdapat banyak algoritma clustering yang dalam penggunaannya tergantung pada tipe data yang akan dikelompokkan dan apa tujuan dari pembuatannya. Dengan menggunakan clustering ini, kita dapat mengklasifikasikan daerah potensial dari pembeli, menemukan pola-pola distribusi secara keseluruhan, dan menemukan keterkaitan yang menarik antara atribut data(Gunawan, Firman, & Faiza,

2016).

6. K-Means

K-Means merupakan suatu algoritma yang digunakan dalam pengelompokkan secara pertisi yang memisahkan data ke dalam kelompok yang berbeda–berda. Algoritma ini mampu meminimalkan jarak antara data ke cluster(Pulungan, Poningsih, & Satria, 2019). Langkah-langkah algoritma K-Means adalah sebagai berikut:

a. Pilih secara acak k buah data sebagai pusat cluster.
b. Jarak antara data dan pusat cluster dihitung menggunakan Euclidian Distance. Untuk menghitung jarak semua data ke setiap titik pusat cluster dapat menggunakan teori jarak Euclidean yang dirumuskan sebagai berikut:

$$D(i,j) = \sqrt{(x_{1i} - x_{1y})^2 + (x_{2i} - x_{2y})^2 + \ldots\ldots\ldots\ldots + (x_{ki} - x_{ky})^2}$$

dimana:

D (i,j)	= Jarak data ke i ke pusat *cluster* j
X_{ki}	= Data ke i pada atribut data ke k
X_{kj}	= Titik pusat ke j pada atribut ke k

c. Data ditempatkan dalam cluster yang terdekat, dihitung dari tengah cluster.
d. Pusat cluster baru akan ditentukan bila semua data telah ditetapkan dalam cluster terdekat.Proses penentuan pusat cluster dan penempatan data dalam cluster diulangi sampai nilai centroid tidak berubah lagi.

7. Euclidean Distance

Dari beberapa penelitian untuk clustering,pengukuran jarak yang pada umumnya sering digunakan adalah jarak Euclidean. Eucledian Distance dianggap sebagai distance matrix yang mengadopsi prinsip Phytagoras. Hal ini dikarenakan pola perhitungannya yang menggunakan aturan pangkat dan akar kuadrat. Eucledian akan memberikan hasil jarak yangrelatif kecil karena menggunakan aturan akar kuadrat (Diwahana Mutiara, Abdul, & Moch. Arief, 2019).

8. Purity Validasi

Purity digunakan untuk menghitung kemurnian dari suatu cluster yang direpresentasikan sebagai anggota cluster yang paling banyak sesuai (cocok) disuatu kelas. Nilai purity yang semakin mendekati 1 menandakan semakin baik cluster yang diperoleh (Burhanuddin, Utami, & Pramono, 2017).

HASIL DAN PEMBAHASAN

1. Deskripsi Data

Data dalam penelitian ini berjumlah sebanyak 550 data customer atau pelanggan Yamaha Nusantara Motor. Preprosessing data digunakan untuk menghilangkan duplikasi data dan menyiapkan data untuk diolah dengan metode tersebut. Setelah melalui preprosessing didapatkan data sebanyak 500.

2. Proses Data Training Menggunakan Metode K-NN dan K-Means

Dalam proses ini data diambil berjumlah 100 data sebagai data training yang akan diolah menggunakan metode K-NN terlebih dahulu untuk menentukan klasifikasi kelayakan sebuah data. Proses metode K-NN menggunakan nilai k sebagai berikut sesuai tabel 1:

Tabel 1. Niilai K Proses K-NN

X1	X2	X3	X4	X5
0,26	1,61	1,58	1,88	1,62

Dapat dilihat tabel 2 melalui proses tersebut dihasilkan 41 data dinyatakan tidak layak dan 59 data dinyatakan layak. Data yang dinyatakan layak akan melalui proses metode K-Means.

Tabel 2. Data Proses K-NN Dinyatakan Layak

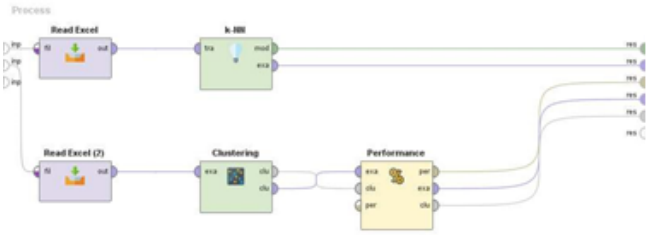
No	ID	X1	X2	X3	X4	X5
1	Customer- 004	0	1	2	1	2
2	Customer- 005	0	1	1	2	1
3	Customer- 006	1	2	2	2	1
4	Customer- 014	1	2	1	2	2
5	Customer- 015	0	1	1	2	2
....						
59	Customer- 099	0	2	1	2	2

3. Proses Data Testing Menggunakan Metode K-NN dan K-Means

Dalam proses ini data diambil berjumlah 400 data sebagai data training yang akan diolah menggunakan metode K-NN terlebih dahulu untuk menentukan klasifikasi kelayakan sebuah data. Proses metode K-NN menggunakan nilai k.

4. Proses Data Training RapidMiner Metode K-NN dan K-Means

Proses data training dalam gambar 2 menggunakan data set sebanyak 100 dilakukan dengan menggunakan RapidMiner dengan 2 metode yaitu metode K-Nearest Neighbor dan metode K-Means

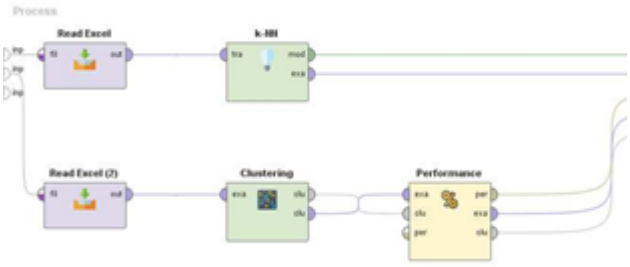


Gambar 2. Proses RapidMiner Data Training K-NN dan K-Means

Data dalam Excel berjumlah 100 data yang dimasukan kedalam fungsi RapidMiner yaitu Read Excel, yang dilanjutkan dengan proses metode K- Nearest Neighbor. Proses tersebut menghasilkan perhitungan K-Nearest Neighbor bahwa data yang dinyatakan layak berjumlah 59 data sedangkan data yang dinyatakan tidak layak berjumlah 41 data sesuai dengan gambar

5. Proses Data Testing RapidMiner Metode K-NN dan K-Means

Proses data training dalam gambar 5 menggunakan data set sebanyak 400 dilakukan dengan menggunakan RapidMiner dengan 2 metode yaitu metode K-Nearest Neighbor dan metode K-Means.



Gambar 3. Proses RapidMiner Data Testing K-NN dan K-Means

Data dalam Excel berjumlah 400 data yang dimasukan kedalam fungsi RapidMiner yaitu Read Excel, yang dilanjutkan dengan proses metode K- Nearest Neighbor. Proses tersebut menghasilkan perhitungan K-Nearest Neighbor bahwa data yang dinyatakan layak berjumlah 210 data sedangkan data yang dinyatakan tidak layak berjumlah 190 data sesuai dengan gambar

6. Purity Validasi Training

Dari tabel 3 purity menunjukan validasi data training dengan data 59 didapatkan nilai purity cluster 1 dengan banyaknya data 21 sebesar 0,3810, nilai purity cluster 2 dengan banyaknya data 6 sebesar 0,6667, nilai purity cluster 3 dengan banyaknya data 13 sebesar 0,0769 dan nilai purity cluster 4 dengan banyaknya data 19 sebesar 0,2105. Nilai purity validasi dari
data training sebesar 1,3351.

Tabel 3. Data Purity Validasi Training

PURITY VALIDASI					
CLUSTER	JUMLAH	t1 (min)	t2 (max)	t3 (average)	Purity
C1	21	8	8	5	0,3810
C2	6	2	4	0	0,6667
C3	13	4	1	8	0,0769
C4	19	6	4	9	0,2105
TOTAL	59				1,3351

7. Purity Validasi Testing

Dari tabel 4 purity menunjukan validasi data training dengan data 210 didapatkan nilai purity cluster 1 dengan banyaknya data 68 sebesar 0,1176, nilai purity cluster 2 dengan banyaknya data 62 sebesar 0,1613, nilai purity cluster 3 dengan banyaknya data 45 sebesar 0,1556 dan nilai purity cluster 4 dengan banyaknya data 35 sebesar 0,2286. Nilai purity validasi dari data training sebesar 0,6631.

Tabel 4. Purity Validasi Testing

PURITY VALIDASI					
CLUSTER	JUMLAH	t1 (min)	t2 (max)	t3 (average)	Purity
C1	68	5	8	7	0,1176
C2	62	7	10	8,54839	0,1613
C3	45	5	7	6,644444444	0,1556
C4	35	6	8	7,085714	0,2286
TOTAL	210				0,6631

PENUTUP

Berdasarkan seluruh hasil tahapan penelitian dan pembahasan, diperoleh kesimpulan sebagai berikut :

1. Dalam metode K-Nearest Neighbor digunakan untuk menghasilkan atau menentukan kelayakan sebuah data dalam proses training yang dihasilkan 59 data dan proses testing yang dihasilkan 210 data yang dinyatakan layak.
2. Dalam metode K-Means digunakan untuk menentukan sebuah cluster yang akan digunakan dalam proses data training dan proses data testing dengan menghasilkan empat buah cluster.
3. Dalam pengujian Purity semua *cluster* dalam penelitian ini diperoleh hasil data training sebesar 1,3351 sedangkan dalam data testing didapatkan nilai sebesar 0,6631, hasil dari proses validasi purity dikatakan baik jika nilai tersebut mendekati nilai 1.

DAFTAR PUSTAKA

Annur, H. 2018. *Klasifikasi Masyarakat Miskin Menggunakan Metode Naïve Bayes*. Ilkom Jurnal Ilmiah, 10(2), 160–165.

Burhanuddin, A., Utami, E., & Pramono, E.2017. *Perbandingan Metode Single Linkage dan Fuzzy C Means Untuk Pengelompokkan Trafik Internet*. Jurnal Teknologi Informasi, XII, 1–6.

Diwahana Mutiara, C., Abdul, S., & Moch. Arief, S. 2019. *Penentuan Prioritas Penerima Dana Bantuan Operasional Pendidikan Lembaga Pendidikan Anak Usia Dini dengan Metode*. 15, 77–92.

Fajrin, A. A., & Maulana, A. 2018. *Penerapan Data Mining Untuk Analisis Pola Pembelian Konsumen Dengan Algoritma Fp-Growth Pada Data Transaksi Penjualan Spare Part Motor*. Kumpulan Jurnal Ilmu Komputer, 5(1).

Gunawan, A., Firman, A. P., & Faiza, R.2016. *Penerapan Data Mining Pemakaian Air Pelanggan Untuk Menentukan Klasifikasi Potensi Pemakaian Air Pelanggan Baru Di PDAM Tirta Raharja Menggunakan Algoritma K-Means.* 2016 (Sentika), 18–19.

Mustofa, Z., & Suasana, I. S. 2018. *Algoritma Clustering K-Medoids Pada E- Government Bidang Information And Communication Technology Dalam Penentuan Status EDGI*. Jurnal Teknologi Informasi Dan Komunikasi, 9, 1–

Nuari, R., Apriliyani, A., Kusrini, & Juwari.2018. *Implementasi Metode K-Nearest Neighbor (Knn) Untuk Memprediksi Varietas Padi Yang Cocok Untuk Lahan Pertanian*. Jurnal Informa Politeknik Indonusa Surakarta 4, 2–8.

Pulungan, W., Poningsih, & Satria, H. 2019. *Pengelompokkan Pada Kendaraan Bermotor Menurut Kegunaannya Menggunakan Metode Data Mining K-Means*. Konferensi Nasional Teknologi Informasi Dan Komputer, 3(1), 746–752.

Putra, N. A., Putri, A. T., Prabowo, D. A., Surtiningsih, L., Arniantya, R., & Cholissodin, I. 2017. *Klasifikasi Sepeda Motor Berdasarkan Karakteristik Konsumen Dengan Metode K-Nearest Neighbour Pada Big Data Menggunakan Hadoop Single Node Cluster*. Jurnal Teknologi Informasi Dan Ilmu Komputer, 4(2), 81

Setiawan, R. 2016. *Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Untuk Menentukan Strategi Promosi Mahasiswa Baru (Studi Kasus : Politeknik Lp3i Jakarta)*. Jurnal Lentera Ict, 3(1), 76–92.

Tino, M., Hedy, I., & Pria, S. 2020. *Aplikasi Seleksi Pendukung Keputusan Dalam Proses Penerimaan Karyawan Menggunakan Metode Weight Product.* INFOTECH: Jurnal Informatika & Teknologi, 1, 29–42.