DATAMINING PADA PENJUALAN AIR BERSIH DI SPAM AKIDAH MENGGUNAKAN ALGORITMA K-MEANS CLUSTERING MENGGUNAKAN RAPIDMINER

Putri Dwi Lestari, Mulyawan
Program Studi Komputerisasi Akuntansi D3
Sekolah Tinggi Manajemen Informatika dan Komputer IKMI Cirebon, Indonesia
lestariputri62@gmail.com

ABSTRAK

Perkembangan zaman membawa perubahan terhadap cara pendistribusian air bersih untuk masyarakat. Di Desa Kecomberan Kecamatan Talun, pendistribusian air bersih dinaungi oleh pemerintah desa dengan diadakannya Sistem Penyedia Air Minum atau SPAM. Setiap bulan, SPAM Akidah akan mencatat pendistribusian air yang digunakan oleh masyarakat dalam buku kubikasi penjualan air bulanan lalu kubikasi tersebut akan dibukukan dalam arsip administrasi milik SPAM. Belum terbacanya informasi mengenai banyaknya jumlah pengguna berdasarkan kategori pengguna hemat dan boros secara sistematis sehingga diperlukan data baru yang menampilkan kelas pelanggan dengan kategori pengguna hemat dan boros. Dengan data set kubikasi periode Januari sampai dengan Desember 2022 milik 241 pelanggan SPAM Akidah Desa Kecomberan diharapkan dapat memberikan informasi baru terkait kelas pelanggan. Pengelompokkan kelas pelanggan menggunakan bantuan datamining dengan algoritma K-Means Clustering. K-Means Clustering dipilih menjadi metode untuk menemukan pola jarak dari data kubikasi ke titik pusat dengan menerapkan perhitungan Euclidien Distance. Perhitungan Euclidien Distance menghasilkan nilai-nilai hasil iterasi akhir dan kelas-kelas pelanggan. Pengujian yang dilakukan memperoleh nilai Davies Bouldin Index (DBI) sebesar 0,235dengan kelas pelanggan irit pada Cluster 0 sebanyak 190 pelanggan dan kelas pelanggan boros pada Cluster 1sebanyak 51 pelanggan.

Kata kunci: *Data mining, Metode Clustering, Algoritma K-Means Clustering, Penjualan kubikasi air, Rapidminer*

1. PENDAHULUAN

Ketersediaan air bersih tidak dapat dipisahkan dengan kehidupan manusia. Perkembangan zaman membawa perubahan dibidang pendistribusian air. SPAM Akidah adalah hasil dari perkembangan zaman dalam bidang pendistribusian air bersih di Desa Kecomberan, Kecamatan Talun.

Budaya masyarakat dan cuaca menjadi faktor yang mempengaruhi penggunaan air bersih dari sumber air ke rumah pelanggan. Pada musim penghujan, seringkali terjadi konsleting mesin sehingga menghambat pengaliran air ke rumah pelanggan. Penggunaan air oleh 241 pelanggan SPAM Akidah akan dicatat setiap bulannya dan dibukukan dalam arsip. Oleh karena itu perlu dikelompokkan pelanggan berdasarkan penggunaan air yang beragam sehingga akan menghasilkan data pelanggan irit dan boros berdasarkan penggunaannya tiap bulan

Riset sebelumnya membahas tentang Penerapan Data Mining menggunakan algoritma K-Means pengelompokkan pelanggan berdasarkan air terjual menggunakan WEKA. Seperti yang dilakukan oleh Pangestu.[1]

Metode yang digunakan adalah *K-Means Clustering* dengan objek penelitian PAM Kerta Raharja dan aplikasi WEKA yang digunakan untuk membantu mengelola data penelitian yang ada. Pada riset tersebut menghasilkan tiga kelompok pelanggan dengan jumlah terkait berdasarkan total pelanggan dan jumlah kubikasi dalam tiga bulan dengan variable nama, jumlah pemakaian awal, jumlah pemakaian akhir dan jumlah pemakaian sekarang.

Data set yang digunakan bersumber dari kubikasi bulanan dalam periode satu tahun penuh. Data ini terdiri dari sebagai berikut:

1. Memiliki 16 atribut yaitu, Nama Pengguna, RT, RW, Blok, Januari, Februari, Maret, April, Mei, Juni, Juli, Agustus, September, Oktober, November, dan Desember.
2. Jumlah pengguna terdiri dari 241 pengguna

Fokus permasalahan ini untuk mengelompokkan kumpulan data kubikasi bulanan pada arsip SPAM Akidah yang belum terorganisir dengan baik dan belum tersusun secara sistematis sehingga menyebabkan ketidaktahuan kepala pimpinan dan masyarakat umum mengenai siapa saja yang termasuk pelanggan yang irit dan boros dengan menerapkan algoritma *K-Means Clustering* dan juga dengan bantuan aplikasi data mining yaitu *Rapidminner.*

Adapun yang menjadi tujuan kali ini adalah untuk memberikan informasi mengenai data pelanggan kubikasi baru yang lebih tersusun dan sistematis. Sehingga data tersebut bisa dipahami dengan mudah.

2. TINJAUAN PUSTAKA

2.1. *Data mining*

Data mining adalah proses untuk menganalisa data dalam jumlah banyak dan menghubungkannya dengan data lain sehingga mendapatkan data yang tersusum secara sistematis. Data yang sudah diperoleh akan menghasilkan sebuah pola

Dalam data mining, perlu adanya KDD (Knowledge Discovery in Database Process). Proses dalam KDD terdiri dari tahap pembersihan data, integrasi data, pemilihan data, transformasi data, evaluasi pola, dan presentasi pengetahuan. Penggunaan data mining terdapat pada data relasi, data transaksi, multimedia, web, dan text.

2.2. *Clustering*

Dalam jurnal dengan judul Penerapan Data Mining Pengelompokan Data Pengguna Air Bersih Berdasarkan Keluhannya Menggunakan Metode Clustering Pada PDAM Langkat yang disusun oleh Karin Annisa, Budi Serasi Ginting, Mili Alfhi Syari. Clustering merupakan proses pengelompokan satu set objek data ke dalam himpunan bagian yang disebut dengan cluster[2]

Metode dalam clustering dibagi menjadi 2, yaitu:

1. Hierarchical clustering, secara sederhana metode ini digunakan untuk mengelompokkan objek dengan kemiripan yang sangat mirip dan yang sangat tidak mirip[3].
2. Non-hierarchical clustering, secara sederhana, metode ini digunakan untuk mengelompokkan objek dengan karakteristik berbeda sehingga data tersebut memiliki atribut yang lebih sedikit. Algoritma *K-Means Clustering* menggunakan metode *clustering* ini

2.3. *Algoritma K-Means Clustering*

K-Means Clustering adalah sebuah metode yang ada dalam data mining dengan proses awal mengambil sebagian komponen populasi untuk dijadikan titik pusat di *cluster* awal. Keunggulan *K-Means Clustering* adalah mampu memproses data dengan jumlah banyak dan dengan waktu yang singkat. Dalam penggunaannya, terdapat beberapa aturan yang harus dipahami seperti aturan mengenai jumlah *cluster* yang diperlukan dan atributnya adalah angka (*numerik*)[4].

Adapun proses dari *K-Means Clustering*:

1. Menentukan jumlah K *cluster*. Syarat untuk menentukan nilai K adalah nilainya lebih kecil dari jumlah data.
2. Menentukan titik yang dipilih secara acak sebanyak K buah di awal proses. Dimana titik ini akan menjadi pusat (*Centroid*) dari masing-masing *cluster*.
3. Menghitung jarak dan alokasikan data ke dalam *centroid* sehingga menghasilkan nilai rata-rata.
4. Menentukan *centroid* baru atau nilai rata-rata pada tiap *cluster*.
5. Apabila dalam data ada yang berpindah *cluster* dan adanya perubahan nilai *centroid* yang sudah ditetapkan sebelumnya, maka lakukan lagi perhitungan jarak dalam *centroid*.

2.4. *RapidMiner*

RapidMiner adalah aplikasi yang digunakan untuk mengolah data dan menganalisis data. RapidMiner memberikan solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi.

2.5. Penjualan

Penjualan berawal dari kata jual. Dalam Kamus Besar Bahasa Indonesia (KBBI), Penjualan adalah proses yang dilakukan oleh penjual dalam upaya untuk memenuhi kebutuhan pembeli. Penjualan bisa bersifat langsung dan tidak langsung

2.6. Air Bersih dan SPAM

Menurut (Ulfah, 2022) dalam jurnal yang berjudul efektivitas pelayanan air bersih program PAMSIMAS. Air bersih adalah air yang aman digunakan untuk air minum dan pemakaian-pemakaian lain karena telah bersih dari bibit-bibit penyakit yang dapat membahayakan kesehatan[5]

SPAM atau Sistem Penyedia Air Minum adalah salah satu bentuk sektor publik yang merupakan bagian dari perekonomian nasional yang dinaungi oleh pemerintah desa[6].

3. METODE PERANCANGAN

Pada perancangan ini memuat beberapa tahapan seperti tahapan analisa dan evaluasi saat survei lapangan, pengumpulan data pendukung, proses KDD, Penerapan *K-Means Clustering* menggunakan *Microsoft Excel* dan tahapan terakhir yaitu penerapan *K-Means Clustering* menggunakan *Rapidminer.*

Gambar 1.Tahapan Perancangan

3.1. Analisa dan Evaluasi (survei) Lapangan.

Tahapan pertama dalam perancangan ini adalah Analisa dan evaluasi lapangan untuk mendapatkan permasalahan yang akan dicarikan solusi yang tepat.

3.2. Pengumpulan Data Pendukung.

Tahapan yang kedua dalam perancangan ini adalah pengumpulan data pendukung. Data yang diperlukan adalah data set dan jurnal-jurnal pendukung perancangan sebagai referensi tambahan.

3.3. Proses KDD (*Knowledge Discovery in Database*)

Tahapan ketiga yang dilakukan dalam perancangan tugas akhir adalah proses KDD[7]. Tahapan KDD adalah sebagai berikut:

1. Seleksi data (Data Selection)
   Seleksi ini dilakukan untuk mendapatkan kelompok data berdasarkan atribut yang diperlukan seperti Nama Pengguna, Alamat, dsb. Dalam perancangan tugas akhi ini terjadi tahap seleksi data dari data kubikasi penjualan air bersih. Dari total 16 atribut data, hanya akan diseleksi menjadi 4 atribut data sebagai fokus pengerjaan perancangan
2. Preprocessing
   Dalam proses preprocessing ini, terdapat tahapan cleaning dan reduction pada data-data yang tidak relevan akan dibersihkan agar data dapat digunakan pada proses berikutnya dan menganti atribut menjadi relevan. Dalam perancangan tugas akhir ini terjadi tahapan cleaning data. Dimana atribut dua belas bulan dihapuskan dan diubah menjadi atribut data baru melalui tahapan reduksi data.
3. Integrasi data (Data Integration)
   Integrasi data adalah penghubungan kembali data yang terpecah. Proses ini dilakukan apabila dalam perancangan tugas akhir menggunakan lebih dari satu sumber data atau menggunakan beberapa periode data sehingga data dari masing-masing sumber atau periode digabungkan menjadi satu kesatuan data. Dikarenakan data yang diterima adalah data yang sudah relevan dimana data perancangan ini menggunakan satu sumber dan satu periode tahun penjualan, maka bisa langsung melakukan proses berikutnya yaitu transformasi data.
4. Transformation data
   Pada tahap ini, kualitas dan hasil data mining akan ditentukan. Atribut data hasil dari cleaning dan mendapatkan proses reduksi akan digunakan sebagai dataset untuk pengujian perancangan tugas akhir ini. Selanjutnya pada proses ini, data penjualan air bersih pada SPAM Akidah yang sudah dirubah menjadi format CSV (Comma Separated Values) akan di proses menggunakan Rapidminner.
5. Data Mining
   Pada proses ini, data akan diproses dan menghasilkan pattern atau pola informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu
6. Evaluasi
   Pada tahapan ini, proses KDD dilakukan, data mining yang dihasilkan adalah alur informasi sehingga dapat ditampilkan dengan tampilan sederhana sehingga mudah untuk di mengerti. Lalu dilakukan presentasi sehingga proses KDD lebih dimengerti

3.4. Penerapan K-Means Clustering pada Microsoft Excel

Tahapan yang keempat adalah penerapan *K-Means Clustering* pada *Microsoft Excel*. Tahapan ini meliputi pengumpulan data kubikasi penjualan air bersih pada SPAM Akidah dan setelah data sudah dikumpulkan, tahap selanjutnya menerapkan *algoritma K-Means Clustering* untuk mengelompokkan kelas-kelas pelanggan menurut jumlah kubikasi bulanannya dan menghitung *Euclidean Distance* secara manual dengan bantuan *Microsoft Excel* [8]. Dalam penerapan ini, menggunakan perhitungan *Euclidean Distance* dimana akan menghasilkan selisih jarak antara data satu dengan titik uji *centroid*. Jarak ini yang akan digunakan sebagai dasar penentuan pengelompokkan kelas-kelas pelanggan.

3.5. Penerapan K-Means Clustering menggunakan Rapidminner.

Pada tahapan terakhir setelah dilakukannya analisis permasalahan, pengumpulan data dan penerapan algoritma adalah pengujian hasil *Davies Bouldin Index* (DBI). Pada tahapan ini akan dilakukan pengujian data untuk mendapatkan hasil setelah menggunakan metode yang sudah diterapkan apakah akan mendapatkan hasil yang dapat menjawab masalah pada rumusan masalah atau sebaliknya. *Algoritma K-Means Clustering* akan efektif digunakan apabila dalam pengujian ini menghasilkan nilai *Davies Bouldin Index* (DBI) tidak lebih dari 0,7.[9]

4. HASIL DAN PEMBAHASAN

Penerapan Algoritma K-Means Clustering Menggunakan Microsoft Excel

1. Menentukan jumlah cluster (K)
   Dalam model ini, jumlah *cluster* (K) yang ditentukan sebanyak K=2. K dalam *cluster* ini untuk menunjukkan kategori pelanggan hemat dan boros.
2. Menentukan titik *Centroid* awal.
   Dalam menentukan *centroid* awal, dilakukan dengan cara mengambil data secara *random* (acak) dari baris ke-151 sebagai C0 dan baris ke-95 sebagai C1. Adapun tabel *centroid* awal adalah sebagai berikut:

Tabel1.Centroid Awal

CENTROID	TST	RTRT
C0	52	4,33
C1	6184	515,33

3. Menghitung Jarak data terhadap centroid
   Menentukan jarak ini dilakukan dengan menerapkan perhiungan *Euclidean Distance* dengan bantuan *Microsoft excel*. Penentuan jarak dapat menghasilkan data cluster 0 sebanyak 136 anggota dan data *cluster* 1 sebanyak 105 anggota. Perhitungan jarak ini dilakukan untuk mendapatkan jarak cluster 0 ke *centroid* dan jarak

cluster 1 ke *centroid*, selanjutnya di buat jarak terdekatnya berdasarkan *cluster*. Menghitung jarak titik *centroid* terhadap data adalah menggunakan formula dalam *Microsoft Excel* yaitu:

C0 = SQRT(((C2-G2)^2+(D2-H2)^2))
C1 = SQRT(((C2-G3)^2+(D2-H3)^2))

4. Menentukan Centroid Baru
   Setelah mendapatkan hasil dari masing-masing cluster, selanjutnya adalah tahapan mementukan centroid baru dengan cara mencari nilai rata-rata tiap clusternya. [10]
   Setelah dilakukan sebanyak 5 iterasi dan pada iterasi ke-4 dan ke-5 anggota cluster tidak mengalami perubahan nilai, maka iterasi dianggap selesai dan dilanjutkan dengan tahap berikutnya[2]. Adapun hasil dari tiap iterasi dapat dilihat pada tabel berikut:

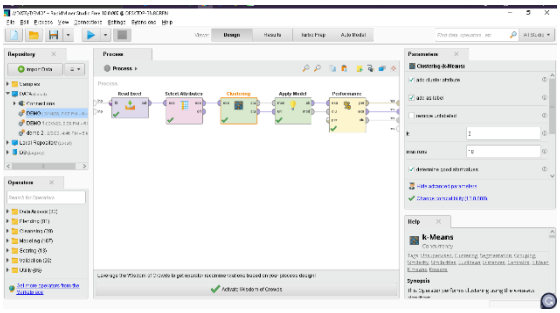
Tabel 2. Hasil Iterasi

Iterasi	Jumlah anggota cluster	
	C0	C1
1	136	105
2	164	77
3	240	1
4	189	52
5	189	52

Jumlah anggota cluster yang mendekati jumlah cluster pada Rapidminner ada pada iterasi ke-4 dan tidak berubah ketika dilakukan iterasi selanjutnya, yaitu 190 anggota C0 dan 51 anggota C1.

4.1. Penerapan Algoritma K-Means Clustering Menggunakan Rapidminer

Pemodelan pada Rapidminer ditampilkan pada gambar berikut



Gambar 2. Model Data Mining

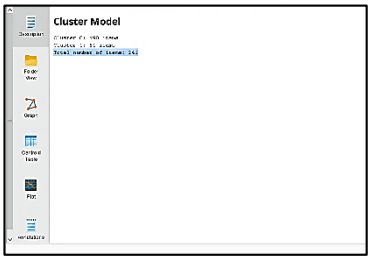
4.2. *Read Excel*

Port pertama pada model data mining adalah *Read excel*. Dimana untuk membaca data set dalam format *excel*. Data set dimasukan dengan klik dua kali pada *port read excel* maka akan masuk pada parameter *import configuration wizard* dan akan menampilkan lokasi data dalam folder pemilik lalu pilih data set yang akan digunakan

4.3. *Select Attribute*

Dalam *parameters select attribute, attribute filter type* diatur menjadi subset dan memilih atribut mana saja yang akan digunakan

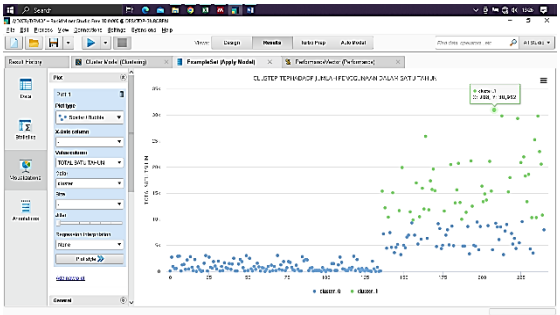
1. *Clustering*
   penambahan port *Clustering K-Means* dipilih dalam proses view untuk mengelompokkan kelas kelas pelanggan dalam data set kubikasi penjualan air bersih di SPAM Akidah berdasarkan rata-rata pengguna dalam satu tahun dan total penggunaan dalam satu tahun. Kelas yang akan dicari sebanyak 2 *cluster*.
2. Performance
   Performance yang digunakan adalah *cluster distance performance*. Lalu klik *run* untuk menampilkan hasil pengolaan data.



Gambar 3. Cluster Model

4.4. *Cluster Model*

Pada Cluster Model menampilkan jumlah anggota C0 sebanyak 190 anggota dan C1 jumlah anggota C1 sebanyak 51 anggota dari keseluruhan data sebanyak 241 anggota



Gambar 4. Visualisasi Anggota Cluster

4.5. Visualisasi Example

Pada gambar 4, menampilkan cluster 0 dan cluster 1 berdasarkan total pengguna satu tahun penggunaan. Titik biru menunjukkan anggota cluster 0 dan titik hijau menunjukkan anggota cluster 1. Berdasarkan total pengguna satu tahun, cluster 1 adalah cluster tertinggi daripada cluster 0.
Penetapan cluster irit dan boros dapat dilihat dengan range pada tabel berikut:

Tabel 3. Tabel Kategoris

Cluster	TST	Rata-rata
0	6 – 9432	0.50 - 786
1	9565 – 30952	797,08 - 2579,33

5. KESIMPULAN DAN SARAN

Penerapan Algoritma K-Means Clustering pada Microsoft Excel dengan menghitung jarak menggunakan Euclidean Distance menampilkan hasil anggota cluster 0 sebanyak 189 anggota, sedangkan cluster 1 sebanyak 52 anggota dengan nilai Davies Bouldin Indeks sebesar 0,989205

Penerapan Algoritma K-Means Clustering pada Rapidminer menampilkan hasil anggota cluster 0 sebanyak 190 anggota, sedangkan cluster 1 sebanyak 51 anggota dengan nilai Davies Bouldin Index sebesar 0,235

Perbedaan hasil ini disebabkan oleh adanya perbedaan proses dalam penerapan Microsoft Excel dan Rapidminer. Dalam Microsoft excel, penetapan nilai centroid dilakukan secara acak dari data uji. Sedangkan dalam Rapidminer, nilai centroid sudah ada pada table centroid. Dan dalam penerapan perhitungan jarak menggunakan Euclidean Distance pada Microsoft Excel mengenal adanya iterasi awal. Sedangkan dalam penerapan perhitungan jarak pada Rapidminer hanya mengenal adanya hasil akhir tanpa menampilkan jumlah iterasinya.

DAFTAR PUSTAKA

[1] A. Pangestu, "PENERAPAN DATA MINING MENGGUNAKAN ALGORITMA K- MEANS PENGELOMPOKAN PELANGGAN BERDASARKAN KUBIKASI AIR TERJUAL MENGGUNAKAN WEKA," *PENERAPAN DATA Min. MENGGUNAKAN Algoritm. K-MEANS PENGELOMPOKAN Pelangg. BERDASARKAN KUBIKASI AIR TERJUAL MENGGUNAKAN WEKA*, vol. 11, no. 3, pp. 67–71, 2021.

[2] K. Annisa, B. S. Ginting, and M. A. Syari, "Penerapan Data Mining Pengelompokan Data Pengguna Air Bersih Berdasarkan Keluhannya Menggunakan Metode Clustering Pada PDAM Langkat," vol. 6341, no. April, 2022.

[3] U. Syafiyah, I. Asrafi, B. Wicaksono, D. P. Puspitasari, and F. M. Sirait, "Analisis Perbandingan Hierarchical dan Non-Hierarchical Clustering Pada Data Indikator Ketenagakerjaan di Jawa Barat Tahun 2020," vol. 2020, pp. 803–812, 2020.

[4] M. safii Sinaga lestari, ahmad abdullah, "PENERAPAN DATA MINING PADA JUMLAH PELANGGAN PERUSAHAAN AIR BERSIH MENURUT PROVINSI MENGGUNAKAN METODE K-MEANS CLUSTERING," *PENERAPAN DATA Min. PADA JUMLAH Pelangg. Perusah. AIR BERSIH MENURUT PROVINSI MENGGUNAKAN Metod. K-MEANS Clust.*, vol. 2, no. 2, pp. 119–125, 2019.

[5] Ulfah, "efektivitas pelayanan air bersih program pamsimas," pp. 1–146, 2022.

[6] S. Husain *et al.*, "Evaluasi Penetapan Tarif Air Minum Pada Perusahaan Daerah Air Minum (PDAM) Kabupaten Kepulauan Sangihe Evaluation of Drinking Water Tariff Determination at The Regional Drinking Water Company (PDAM) of Sangihe Islands Regency," vol. 6, no. 1, pp. 377–388, 2022.

[7] M. Alfariqi, "Implementasi Data Mining Metode Clustering Pada Toko Buku Bi-Obses Bandung Dengan Penerapan Algoritma K-Means," *J. Ilm. Komput. dan Inform. (KOMPUTA).*, no. Implementasi Data Mining Metode Clustering Pada Toko Buku Bi-Obses Bandung Dengan Penerapan Algoritma K-Means, 2016.

[8] Y. Miftahuddin, S. Umaroh, and F. R. Karim, "PERBANDINGAN METODE PERHITUNGAN JARAK EUCLIDEAN , HAVERSINE , (STUDI KASUS : INSTITUT TEKNOLOGI NASIONAL BANDUNG)," vol. 14, no. 2, pp. 69–77, 2020.

[9] W. Gie and D. Jollyta, "Perbandingan Euclidean dan Manhattan Untuk Optimasi Cluster Menggunakan Davies Bouldin Index: Status Covid-19 Wilayah Riau," *Pros. Semin. Nas. Ris. Dan Inf. Sci.*, vol. 2, no. April, pp. 187–191, 2020.

[10] N. Putu, E. Merliana, P. Studi, M. Teknik, F. T. Industri, and U. A. Jaya, "ANALISA PENENTUAN JUMLAH CLUSTER TERBAIK PADA METODE K-MEANS," pp. 978–979, 2017.