PENERAPAN *DATA MINING* UNTUK PENGELOMPOKAN PRODUK PENJUALAN MENGGUNAKAN ALGORITMA *K-MEANS* (STUDI KASUS : TOKO AGUNG MAKMUR JAYA)

Fajar Dwi Agustiar, Betha Nurina Sari, Iqbal Maulana
Informatika, Universitas Singaperbangsa Karawang
Jl. HS. Ronggo Waluyo, Telukjambe Timur, Karawang, Indonesia
fajar.dwi18177@student.unsika.ac.id

ABSTRAK

Perkembangan dan persaingan dalam bisnis perdagangan serta kemajuan teknologi informasi membuat semakin ketatnya persaingan pasar untuk memenuhi tuntutan pelanggan yang terus meningkat. Toko Agung Makmur Jaya masih sering kesulitan dalam menentukan strategi bisnis seperti apa yang paling efektif. Berdasarkan data hasil penjualan selama satu tahun, bulan Juli 2023 – Juni 2024 tren penjualan Toko Agung Makmur Jaya cenderung menurun. Toko Agung Makmur Jaya dapat memanfaatkan fitur laporan penjualan produk setiap bulan yang sudah ada untuk menentukan pola penjualan yang efektif. Penelitian ini bertujuan menerapkan *clustering k-means* untuk pengelompokan produk Toko Agung Makmur Jaya. Penelitian ini dilakukan dengan bahasa pemrograman Python dan menggunakan metodologi CRISP-DM. DBI digunakan untuk menentukan jumlah *cluster* terbaik, yaitu 5 *cluster* dengan nilai DBI 0.453559. Hasil dari penelitian ini berupa pengelompokan produk dari data penjualan Toko Agung Makmur Jaya, pada *Cluster 4* dikategorikan sangat laku dan omzet sangat besar, *Cluster 2* dikategorikan laku dan omzet besar, *Cluster 1* dikategorikan cukup laku dan omzet sangat besar, *Cluster 3* dikategorikan tidak laku dan omzet tidak besar, dan *Cluster 0* dikategorikan sangat tidak laku dan omzet sangat tidak besar. Hasil *clustering k-means* dievaluasi menggunakan *Silhoutte Coeffecient* mendapatkan nilai sebesar 0,68 yang mana nilai tersebut masuk ke dalam kategori struktur klaster standar.

Kata kunci : *Data Mining, Pengelompokan Produk Penjualan, K-Means, DBI, Silhoutte Coeffecient*

1. PENDAHULUAN

Perkembangan dan persaingan dalam bisnis perdagangan serta kemajuan teknologi informasi saling berkaitan, membuat semakin ketatnya persaingan pasar untuk memenuhi tuntutan pelanggan yang terus meningkat. Strategi dan kecerdasan bisnis diperlukan untuk dapat terus memenuhi keinginan pelanggan dan tuntutan pasar [1].

Dalam dunia bisnis yang semakin kompetitif saat ini, orang-orang dituntut untuk terus mengembangkan bisnisnya agar dapat bertahan dalam persaingan terutama dalam hal penjualan, hal ini menuntut para pelaku bisnis untuk menemukan suatu pola-pola tertentu yang bisa meningkatkan penjualan dan pemasaran, salah satunya dengan pemanfaatan data penjualan [2].

Toko Agung Makmur Jaya masih sering kesulitan dalam menentukan strategi bisnis seperti apa yang efektif, hal tersebut berdampak pada hasil penjualan yang tidak stabil dan cenderung menurun. Saat ini Toko Agung Makmur Jaya menjalankan proses bisnis nya menggunakan aplikasi kasir *online* berbasis *website* bernama FolioPOS, pada aplikasi tersebut tidak hanya menyediakan fitur untuk kasir saja, tetapi terdapat fitur manajemen stok, dan juga fitur laporan data penjualan. Namun data penjualan tersebut tidak dimanfaatkan, dan juga Toko Agung Makmur Jaya juga memiliki permasalahan seperti barang yang tidak laku menumpuk di rak penjualan dan melakukan pengadaan stok yang tidak tepat.

Dengan permasalahan yang sudah ada, maka Toko Agung Makmur Jaya perlu memanfaatkan data penjualan yang sudah ada untuk dilakukan analisis data. Diperlukan suatu teknik atau perangkat untuk membantu mentransformasikan data dalam jumlah yang besar tersebut menjadi informasi berguna, yaitu dengan menerapkan *data mining* [3].

Salah satu teknik *data mining* yang bisa menangani permasalahan pada penelitian ini bisa menggunakan *clustering*.

Clustering merupakan suatu proses yang bertujuan untuk mengelompokkan data yang memiliki kemiripan antara satu data dengan data lainnya ke dalam klaster atau kelompok, sehingga data dalam satu klaster memiliki tingkat kemiripan yang maksimum dan data antar klaster memiliki kemiripan yang minimum [4].

Dengan melakukan *clustering* bisa mengetahui pengelompokan barang yang terjual dari data laporan penjualan, seperti mengetahui kelompok barang yang laku sampai barang yang tidak laku, sehingga dari informasi tersebut bermanfaat untuk strategi bisnis.

Penelitian ini akan melakukan penerapan *clustering k-means* pada data penjualan produk di Toko Agung Makmur Jaya, lalu akan dilakukan pemilihan nilai k (*cluster*) yang optimal berdasarkan nilai *Davies-Bouldin Inde*x yang paling minimum untuk menghasilkan *clustering* yang terbaik, selanjutnya dari hasil *clustering* tersebut akan di uji kualitas terhadap struktur *cluster* nya menggunakan *Silhoutte Coeffecient*.

2. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Adapun penelitian terdahulu yang dijadikan sebagai acuan pada penelitian ini. Penelitian yang dilakukan oleh Pangestu, B., dkk, pada penelitian ini menggunakan data penjualan obat pada Apotek Amarta Sehat dalam penerapan *clustering k-means*, hasilnya terdapat 5 *cluster*, terdiri dari jenis obat dengan kategori sangat laris (C4) terdapat 11 obat, jenis obat dengan kategori cukup laris (C3) terdapat 76 obat, jenis obat dengan kategori sedang (C2) terdapat 131 obat, jenis obat dengan kategori kurang laris (C1) terdapat 399 obat dan jenis obat dengan kategori sangat kurang laris (C0) terdapat 326 obat [5].

Penelitian selanjutnya yang dilakukan oleh Muningsih, E., dkk., pada penelitian ini melakukan perhitungan *Davies-Bouldin Index* dalam mencari jumlah *cluster* yang optimal pada algoritma *k-means* untuk klasterisasi provinsi yang memiliki potensi industri, dari skenario 2–10 *cluster*, hasilnya 3 *cluster* sebagai jumlah *cluster* yang optimal dengan nilai DBI sebesar 0,175. Hasil klasterisasi pada penelitian ini, yaitu terdiri dari kelompok provinsi dengan potensi industri sedikit (C0) sebanyak 31 provinsi, kelompok provinsi dengan potensi industri banyak (C1) sebanyak 2 provinsi dan kelompok provinsi dengan potensi industri sedang (C2) sebanyak 1 provinsi [6].

Penelitian terdahulu berikutnya yang dilakukan oleh Rahmawati, E., penelitian ini menggunakan metode *Silhoutte Coeffecient* sebagai evaluasi algoritma *k-means* dalam mengelompokan daerah yang terdampak tanah longsor di provinsi Jawa Tengah tahun 2020–2022. Dua *cluster* dibentuk dalam pada algoritma *k-means*, hasil klasterisasi terdiri dari kelompok daerah kota/kabupaten dengan dampak tinggi (C0) sebanyak 7 provinsi dan kelompok daerah kota/kabupaten dengan dampak rendah (C1) sebanyak 24 kota/kabupaten. Hasilnya nilai *Silhoutte Coeffecient* yang didapat sebesar 0,48 yang artinya struktur *cluster* lemah [7].

2.2. Data

Data adalah fakta dan statistik yang dikumpulkan bersama untuk berbagai jenis analisis atau digunakan sebagai referensi untuk mendukung berbagai jenis penelitian atau pendapat [8].

2.3. *Data Mining*

Data mining merupakan rangkaian proses yang menggali nilai tambah berupa informasi yang selama ini belum diketahui secara manual dari suatu *database*. *Data mining* merupakan metode yang utamanya digunakan untuk mencari pengetahuan yang tersembunyi dalam *database* yang berukuran besar, sehingga sering disebut juga sebagai *knowledge discovery database* (KDD) [9].

2.4. *Clustering*

Clustering adalah metode pengelompokan data ke dalam beberapa *cluster* (kelompok) berdasarkan tingkat kemiripan yang tinggi antar data dalam satu klaster dan tingkat kemiripan yang rendah antar klaster [10].

Clustering juga disebut sebagai *segmentation*. Metode ini digunakan untuk mengidentifikasi kelompok alami dari sebuah kasus yang didasarkan pada sebuah kelompok atribut, sehingga mengelompokan data dalam satu klaster memiliki kemiripan nilai pada suatu atribut [9].

2.5. K-Means

K-Means merupakan algoritma yang mudah untuk diimplementasikan dan memiliki kompleksitas waktu dan ruang yang relatif kecil dalam prosesnya. Algoritma ini juga merupakan algoritma yang cukup efisien dalam komputasinya dan memberikan hasil yang cukup baik. Menurut Hans & Kamber, algoritma *k-means* bekerja dengan membagi data ke dalam jumlah cluster k yang telah ditentukan. Langkah-langkah dasar untuk algoritma *k-means* adalah [8]:

a. Tentukan jumlah klaster (k) yang diinginginkan.
b. Pilih k *centroid* secara acak sebagai awal setiap klaster.
c. Setiap titik data, hitung *Eucludean Distance* antar titik data tersebut ke setiap pusat klaster dengan persamaan sebagai berikut:

$$d\,(p,q) = \sqrt{((x_2 - x_1)^2 + \ldots + (z_n - z_{n-1})^2)} \quad (1)$$

dimana $(x_2, \ldots, z_n)$ adalah koordinat pusat klaster dan $(x_1, \ldots, z_{n-1})$ adalah koordinat titik data.

d. Setiap titik data, hitung *Eucludean Distance* antar titik data tersebut ke setiap pusat klaster dengan persamaan sebagai berikut:

$$v = \frac{\sum_{i=1}^{n} x_i}{n} \quad (2)$$

Keterangan:
v = pusat *cluster*
x_i = titik data ke-i
n = banyaknya titik data dalam satu cluster

e. Ulangi langkah ke-3 dan ke-4 sampai tidak ada perubahan dalam pengelompokan titik data atau hingga mencapai kriteria berhenti yang ditentukan.

Pada langkah ke-3, titik data akan ditempatkan dalam klaster yang memiliki pusat terdekat berdasarkan jarak *Euclidean*. Dalam langkah ke-4 pembaruan pusat klaster, pusat baru dihitung dengan mengambil rata-rata dari semua titik data dalam klaster.

2.5. Davies-Bouldin Index (DBI)

Davies-Bouldin Index (DBI) adalah suatu metrik yang digunakan untuk mengevaluasi atau menilai hasil dari algoritma pengelompokan (*clustering*). Metrik ini pertama kali diperkenalkan oleh David L. Davies dan Donald W. Bouldin pada tahun 1979. Dengan nilai DBI, suatu *cluster* akan dianggap memiliki skema

clustering yang optimal jika memiliki nilai DBI yang minimum (mendekati nol) [11].

2.6. Silhoutte Coeffecient

Silhouette Coefficient digunakan untuk mengevaluasi kualitas dan kekuatan klaster, yang mengukur sejauh mana objek ditempatkan dengan baik di dalam suatu klaster. Metode ini menggabungkan aspek-aspek separasi dan kohesi [12].

Tabel 1. Kriteria Nilai Silhoutte Coeffecient

Nilai Silhoutte Coeffecient	Kriteria
$0,7 < SC \leq 1,0$	Struktur *cluster* kuat
$0,5 < SC \leq 0,7$	Struktur *cluster* standar
$0,25 < SC \leq 0,5$	Struktur *cluster* lemah
$SC \leq 0,5$	Tidak memiliki struktur

Tabel 1 merupakan kriteria pengukuran nilai *Silhoutte Coeffecient* menurut Rousseeuw (1987) [13]. Rentang nilai *Silhoutte Coeffecient* dari 0–1, jika nilainya mendekati 0 maka kulitas suatu klaster semakin tidak baik, tetapi jika nilainya mendekati 1 maka kualitas suatu klaster semakin baik.

2.7. Pemrograman Python

Python adalah bahasa pemrograman dengan model paradigma pemrograman orientasi objek. Pemrograman Python ini biasa digunakan untuk pengembangan perangkat lunak dan dapat dijalankan melalui berbagai sistem operasi. Saat ini, Python salah satu bahasa yang populer pada bidang data science dan data analyst. Hal ini dikarenakan terdapat library – library yang didalamnya menyediakan fungsi analisis data dan fungsi *machine learning*, *data preprocessing tools*, serta visualisasi data pada pemrograman Python [14].

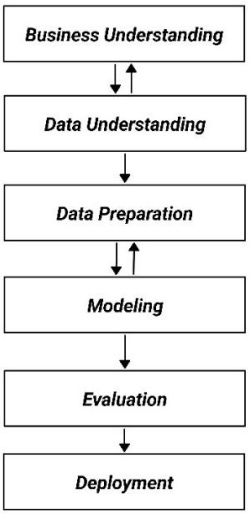
2.8. Google Colaboratory

Colaboratory, atau "Colab" adalah produk dari *Google Research* yang memungkinkan pengguna untuk menulis dan menjalankan kode Python secara bebas melalui peramban *web*, sangat cocok untuk *machine larning*, analisis data, pendidikan [15].

Google Colaboratory yang dibangun dengan menggunakan elemen dari Jupyter serta sebagian besar *library* tersedia untuk diterapkan ke dalam sebuah penelitian salah satunya adalah *data mining* [16].

3. METODE PENELITIAN

Pada penelitian ini akan menggunakan menggunakan metodologi CRISP-DM (*Cross Industry Standard Process for Data Minig*) sebagai tahapan proses *data mining*, terdiri dari fase *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment*. Adapun detail tahapan yang akan dilakukan seperti pada Gambar 1 sebagai berikut.



Gambar 1. Metode Penelitian

3.1. *Business Understanding*

Pada fase *Business Understanding* akan dilakukan tahapan mengeidentifikasi tujuan bisnis dengan mengetahui masalah yang ingin diselesaikan, menilai situasi, dan menentukan tujuan *data mining*.

3.2. *Data Understanding*

Pada fase *Data Understanding* akan dilakukan tahapan memilih data yang sesuai dengan tujuan *data mining* dan mendeskripsikan data tersebut untuk memahami data yang telah didapat.

3.3. *Data Preparation*

Pada fase *Data Preparation* akan dilakukan tahapan seleksi atribut dari data yang telah didapat. Data yang didapat akan dilakukan pembersihan dengan melakukan penanganan duplikasi data, *missing value*, dan *outlier*.

3.4. *Modeling*

Pada fase *Modeling* akan dilakukan tahapan membangun model *clustering* menggunakan algoritma *k-means*, dalam menerapkan *k-means* akan menggunakan nilai DBI (*Davies-Bouldin Index*) untuk mencari jumlah k *cluster* yang optimal. Lalu model *clustering k-means* akan dibentuk dengan jumlah *cluster* berdasarkan nilai DBI yang optimal, yaitu nilai DBI yang paling minimum.

3.5. *Evaluation*

Pada fase *Evaluation* akan dilakukan proses uji kulitas terhadap model *clustering k-means* yang telah dilakukan menggunakan *Silhoutte Coeffecient*. *Silhoutte Coeffecient* digunakan untuk mengetahui apakah model *clustering k-means* yang telah dibuat memiliki struktur yang kuat atau tidak.

3.6. *Deployment*

Pada fase *Deployment* akan dilakukan pembuatan laporan mengenai klasterisasi produk penjualan Toko Agung Makmur Jaya, dimana setiap klaster memiliki

intepretasi, sehingga dari klasterisasi produk tersebut bisa sebagai informasi yang berguna untuk strategi bisnis bagi pemilik Toko Agung Makmur Jaya.

4. HASIL DAN PEMBAHASAN

4.1. *Business Understanding*

Pada fase ini dilakukan identifikasi tujuan bisnis yang ingin dicapai dengan mengetahui permasalahan yang ingin ditangani, adapun masalah yang ingin ditangani oleh Toko Agung Makmur Jaya, yaitu banyak barang yang menumpuk di rak penjualan dan pengadaan stok yang tidak tepat, dari masalah tersebut maka tujuan bisnis yang ingin dicapai, yaitu mengetahui informasi pengelompokan produk untuk strategi penjualan.

Adapun fakta situasi pada Toko Agung Makmur Jaya, bahwa toko tersebut sudah menggunakan layanan kasir *online* FolioPOS untuk menjalankan proses penjualanannya, dan terdapat data laporan penjualan yang tidak dimanfaatkan.

Setelah mengetahui tujuan bisnis yang ingin dicapai dan mengetahui situasi fakta Toko Agung Makmur Jaya, maka dapat ditentuan tujuan *data mining* pada penelitian ini adalah *clustering* produk penjualan Toko Agung Makmur Jaya dari data laporan penjualan.

4.2. *Data Understanding*

Pada fase *Data Understanding* akan dilakukan tiga tahap, yaitu tahap pengumpulan data, tahap pendeskripsian data (*describe data*), dan tahap verifikasi kualitas data (*verify data quality*). Tahap pertama yang dilakukan pada fase *Data Understanding*, yaitu pengumpulan data awal (*collect initial data*). Data didapat dengan mengunduh dari aplikasi kasir *online* FolioPOS Toko Agung Makmur Jaya, data yang digunakan adalah data penjualan dari bulan April–Juni 2024. Adapun *dataset* yanng didapat bisa dilihat pada Tabel 2 sebagai berikut.

Tabel 2. Data penjualan produk bulan April–Juni 2024

Produk	Barcode	Kategori	Jumlah	Penjualan	Diskon	Pajak	Penjualan Bersih	HPP	Laba Kotor
CANADA BOX 8685 FRIEND	null	FRIEND	42	3990000	-69600	0	4059600	3486000	573600
BGL-2 MILLI MEASURIN G CUP 1 LTR	null	BASIC HOME	349	1954400	-45145	0	1999545	1765242	234303
SC-289-1 KOTAK SABUN	null	GOLEDEN HEN	110	198000	-11460	0	209460	173140	36320
M 051 ROSE WATER TUB	null	LUCKY BIRD	118	411500	-42580	0	454080	368750	85330
…	…	…	…	…	…	…	…	…	…
MEJA CHESS SL	8994070008375	SUN LIFE	12	459600	4596	0	455004	414000	41004

Tabel 2 merupakan *dataset* awal yang didapat dari pengunduhan data laporan penjualan produk Toko Agung Makmur Jaya, *dataset* awal yang didapat terdiri dari 10 atribut dan 1680 baris.

Lalu tahap berikut setelah mendapatkan *datset* awal, melakukan deskripsi data. Deskirpsi data dilakukan untuk pemahaman data yang akan digunakan untuk membangun model *clustering*.

Tabel 3. Deskripsi *dataset* awal

Atribut	Deskripsi	Tipe Data
Produk	Nama produk yang terjual	Nominal
Barcode	Angka *barcode* dari produk tertentu untuk keperluan sistem kasir	Nominal
Kategori	Nama perusahaan/pihak pemasok dari produk	Nominal
Jumlah	Jumlah produk yang terjual	Numerik
Penjualan	Total dari harga jual produk yang telah terjual	Numerik
Diskon	Pengurangan dari harga penjualan produk yang telah terjual	Numerik
Pajak	Jumlah pajak dari harga jual produk yang telah terjual	Numerik
Penjualan Bersih	Total dari penjualan bersih dari produk yang terjual	Numerik
HPP	Total dari harga modal produk yang terjual	Numerik
Laba Kotor	Total dari keuntungan produk yang telah terjual	Numerik

Tabel 3 merupakan hasil pendeskripsian data setiap atribut dari *dataset* yang didapat. Dilakukan untuk memahami isi data dari setiap atributnya.

Setelah dilakukan pendeskripsian data, tahap selanjutnya verifikasi kualitas data (*verify data quality*) yang bertujuan untuk melihat *dataset* apakah terdapat *missing value*, *duplicated data*, dan *outlier*. Adapun hasil pengecekan *missing value* dapat dilihat pada Gambar 2 sebagai berikut.

```
datasets.isna().sum()

Produk              0
Barcode          1537
Kategori            1
Jumlah              0
Penjualan           0
Diskon              0
Pajak               0
Penjualan Bersih    0
HPP                 0
Laba Kotor          0
dtype: int64
```

Gambar 2. Pengecekan *missing value*

Gambar 2 merupakan hasil pengecekan *missing value*. Berdasarkan gambar di atas, atribut Barcode dan atribut Katgegori terdapat *missing value*. Karena pada atribut Barcode terdapat 1537 *missing value*, maka atribut tersebut akan dihapus dari *dataset*, sedangkan atribut Kategori terdapat 1 *missing value*, maka produk yang tidak memiliki kategori akan dihapus.

Proses berikutnya, yaitu pengecekan *duplicated data* terhadap atribut Produk, karena data pada atribut tersebut merupakan objek *clustering*, sehingga tidak boleh terjadi duplikasi data.

```
datasets.duplicated("Produk").sum()

57
```

Gambar 3. Pengecekan duplikasi data pada atribut Produk

Gambar 3 menunjukan terdapat 57 duplikasi data pada atribut Produk, maka akan dilakukan agregasi data pada fase berikutnya.

Proses berikutnya adalah pengecekan data *outlier* pada *dataset*, karena data *outlier* akan mempengaruhi hasil dari *clustering k-means*. Pada penelitian ini, data *outlier* sudah diketahui, pada *dataset* yang didapat, kriteria bahwa data tersebut dikatakan *outlier* ketika laba kotor memiliki nilai 0 (nol).

	Produk	Jumlah	Penjualan	Penjualan Bersih	HPP	Laba Kotor
281	CANGKIR SONYA 355 DX CORNELIUS	1	2100	1750	1750	0
446	FH - 3 POT GANTUNG GARDENIA NO. 8	2	23000	20790	20790	0
591	HS - 11 BOTANY WATERING CAN 3 LTR	1	24900	22816	22816	0
775	L 141 WT WATER JUG 8 LT	36	702000	622476	622476	0
798	LUNCH BOX TASMANIA 443 DX CORNELIUS	1	3400	2833	2833	0
1181	PT - 61 TOPLES SALSA POT 970 ML	2	34200	30916	30916	0
1235	S - 807 GARPU MAKAN (24 PCS)	72	626400	559440	559440	0
1236	S - 808 SENDOK MAKAN (24 PCS)	58	504600	450660	450660	0
1341	SEROK PARABOLA GAGANG KAYU 30 NR	1	34000	30000	30000	0
1411	SW - 2289 SEALWARE 6 LITER	29	536500	481052	481052	0
1513	TOPLES 10 LTR 2910 FRIEND	64	985600	853312	853312	0
1516	TOPLES 16 LTR 8116 FRIEND	37	721500	625892	625892	0
1521	TOPLES 3228 FRIEND	21	60900	52500	52500	0
1525	TOPLES 5 LTR 8805 FRIEND	67	495800	427125	427125	0
1528	TOPLES CT 150	4	90000	84000	84000	0

Gambar 4. Produk yang memiliki nilai *outlier* pada atribut Laba Kotor

Gambar 4 merupakan hasil pengecekan *outlier*, terdapat 15 produk yang memiliki nilai *outlier* pada atribut Laba Kotor nya, maka pada fase berikutnya akan dilakukan pengahapusan produk yang memiliki laba kotor 0 (Nol Rupiah). Adapun selanjutnya, kriteria data tersebut dikatakan *outlier* pada kasus ini adalah jika atribut Penjualan Besih memiliki nilai yang sama dengan atribut Jumlah.

	Produk	Jumlah	Penjualan	Penjualan Bersih	HPP	Laba Kotor
264	C - 17 KOTAK SAMPAH LUNABIN 5,3 LTR	21	543900	497202	497574	-372
509	GELAS SET BW 46-13 PORTO CITINOVA MMS	5	132500	77411	122500	-45089
780	L 8030 TEMPAT ES BUAH MILA	172	12900000	11677210	11902400	-225190
1266	SARINGAN CEKO 20 SM	2	20000	2	18800	-18798

Gambar 5. Produk yang memiliki nilai *outlier* pada atribut Penjualan Bersih

Pada Gambar 5 terdapat produk yang memiliki nilai Penjualan Bersih 2 (Dua Rupiah), nilai tersebut merupakan *outlier*, karena nilanya sama dengan atribut Jumlah yang memiliki nilai 2 (dua Pcs), maka produk tersebut akan dihapus pada fase berikutnya.

4.3. *Data Preparation*

Pada fase *Data Preparation* akan dilakukan persiapan data dengan mengolah datanya terlebih dahulu, lalu kemudian data yang sudah diolah digunakan untuk membangun model *clustering* k-*means*. Tahap pertama yang dilakukan pada fase *Data Preparaion* ini adalah tahap *select data* (pemilihan data). Dari *dataset* yang sudah ada, tidak semua atribut digunakan, hanya 6 atribut saja yang digunakan pada proses *Modeling*.

	Produk	Jumlah	Penjualan	Penjualan Bersih	HPP	Laba Kotor
0	HANGER ROTAN SWEET	1182	2393550	2383648	2175176	208472
1	GELAS CHS-9 ZF BBC MMS	966	1408750	1423800	1241310	182490
2	MANGKOK PM 4 MMS	786	1637500	1834000	1473750	360250
3	GARPU MAKAN WATER MARK NK	750	1968750	1991100	1771188	219913
4	GELAS UBM 18 PINGGANG POLOS HG	414	1794000	1830980	1610115	220865
...	...	...	...	...	...	...
1675	SAPU SAWANG RAKBOL DRAGON	1	26400	35000	24416	10584
1676	PT - 86 TOPLES RUBY POT 1100	1	11300	17000	10193	6807
1677	RAK NAGA MAS S/4 DRAGON	1	120000	140000	111000	29000
1678	SIKAT WC BONITA DRAGON	1	4100	4100	3750	350
1679	RAK SUDUT DRAGON	1	38900	38900	36000	2900

1680 rows × 6 columns

Gambar 6. Hasil proses *select data*

Gambar 6 merupakan *dataset* yang sudah dilakukan proses *select data*. Atribut yang dihilangkan adalah atribut Barcode, Kategori, Diskon, dan Pajak. Karena 4 atribut tersebut tidak relevan dijadikan variabel untuk proses *clustering k-means*.

Tahap berikut pembersihan data (*clean data*). Tahap ini dilakukan untuk membesihkan data dari *missing values*, *dulicated data*, dan *outlier*.

```
[107] datasets.drop(datasets[datasets['Produk'] == 'KARDUS'].index,
                    inplace = True)
```

Gambar 7. Proses penghapusan *missing value* atribut Kategori

Gambar 7 merupakan proses penghapusan produk KARDUS, karena produk tersebut tidak memiliki kategori.

```
aggregate_column = ['Jumlah', 'Penjualan', 'Penjualan Bersih', 'HPP', 'Laba Kotor']

datasets = (datasets[aggregate_column].groupby(datasets['Produk']).sum().reset_index())
```

Gambar 8. Proses agregasi data

Gambar 8 merupakan proses penggabungan data terhadap produk yang mengalami duplikasi.

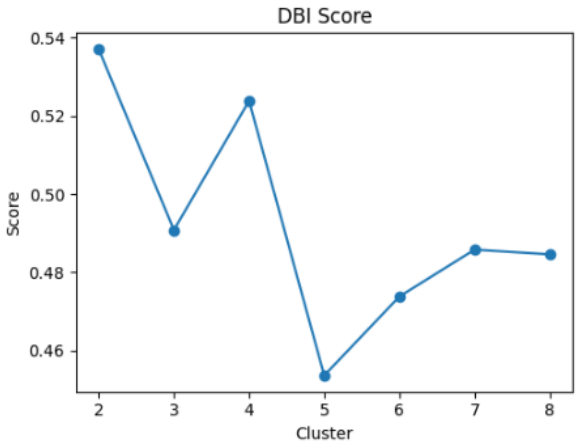
```
detele_produk_index = [1266, 281, 446, 591, 775, 798, 1181, 1235,
                       1236, 1341, 1411, 1513, 1516, 1521, 1525, 1528]

datasets.drop(detele_produk_index, inplace = True)
```

Gambar 9. Proses penghapusan baris data yang terdapat *outlier*

Gambar 9 merupakan proses penghapusan produk yang memiliki data *outlier*, terdapat total 16 produk yang dihapus dari *dataset*, karena terdapat nilai data yang merupakan *outlier* pada atributnya.

4.4. *Modeling*

Setelah *dataset* sudah diolah pada fase *Data Preparation*, fase selanjutnya *Modeling*. Pada fase *Modeling* akan dilakukan proses membangun model (*build model*), model yang dibangun adalah *clustering* dengan algoritma *k-means*. Sebelum menerapkan algoritma *k-means*, perlu menentukan jumlah *cluster* yang ingin dibentuk. DBI (*Davies-Bouldin Index*) diperlukan untuk membantu mencari jumlah *cluster* yang optimal untuk diterapkan pada *dataset* yang sudah ada.



Gambar 10. *Score* DBI pada setiap jumlah *cluster*

Gambar 10 merupakan grafik hasil proses DBI untuk skenario jumlah 2 sampai 8 *cluster*. Jumlah *cluster* dikatakan optimal jika memiliki nilai DBI mendekati 0 (nol) atau yang paling minimum, berdasarkan Gambar 9 diatas, skenario 5 *cluster* memiliki nilai DBI yang paling minimum, yaitu sebesar 0,453559.

Setelah mengetahui bahwa 5 *cluster* merupakan jumlah *cluster* yang optimal, selanjutnya menerapkan algoritma *k-means* dengan nilai k=5.

```
[10] n_clusters = 5
     kmeans = KMeans(n_clusters=n_clusters,
                     random_state=0,
                     n_init="auto").fit(X)
     labels = kmeans.labels_
```

Gambar 11. Proses k-means dengan 5 *cluster*

Gambar 11 merupakan proses menerapakan algoritma *k-means* dengan jumlah 5 *cluster*.

	Produk	Jumlah	Penjualan	Penjualan Bersih	HPP	Laba Kotor	Cluster
0	001ORI RAK SEGI EMPAT SUSUN 3 WS	30	651000	669211	579990	89221	0
1	001PRS RAK SEGI EMPAT SUSUN 3 WS	45	765000	804457	682470	121987	0
2	002ORI RAK OVAL SUSUN 3 WS	10	211000	209101	188330	20771	0
3	002PRS RAK OVAL SUSUN 3 WS	12	204000	209497	181992	27505	0
4	003ORI RAK SEGI TIGA SUSUN 3 WS	37	788100	799620	703296	96324	0
...	...	...	...	...	...	...	...
1601	WIPER AIR DRAT	281	4074500	4280040	3688125	591915	3
1602	WJ - 2286 TEKO AIR 900 ML	9	71100	75350	63630	11720	0
1603	WJ - 2288 TEKO AIR 1500 ML	8	96000	112400	86400	26000	0
1604	WS - 901 WATER SCOOP	184	754400	814311	669392	144919	0
1605	WS - 911 GAYUNG BESAR	23	149500	163220	134366	28854	0

1606 rows × 7 columns

Gambar 12. Hasil proses pelabelan *cluster* terhadap produk

Gambar 12 merupakan hasil proses pelabelan *cluster* terhadap produk. Pada proses ini produk sudah bisa dikelompokan berdasarkan *cluster*.

Setelah produk sudah dilabeli berdasarkan *cluster*, maka selanjutnya proses visualisasi data. Namun karena pada proses *clustering* menggunakan 5 variabel, yaitu jumlah, penjualan, penjualan bersih, HPP, laba kotor, mengakibatkan sulitnya memvisualisasikan data hasil *clustering* menggunakan *scatter plot*. Oleh karena itu dibutuhkan proses reduksi dimensi menggunakan metode PCA (*Principal Component Analysis*), menjadi 2 dimensi.

	PC1	PC2	Cluster
0	-0.663653	-0.116977	0
1	-0.499952	-0.091566	0
2	-1.100474	-0.079071	0
3	-1.085541	-0.072562	0
4	-0.559477	-0.117329	0
...	...	...	...
1601	2.831263	0.024274	3
1602	-1.201431	-0.050281	0
1603	-1.155064	-0.068076	0
1604	-0.246218	0.521426	0
1605	-1.096012	-0.013203	0

1606 rows × 3 columns

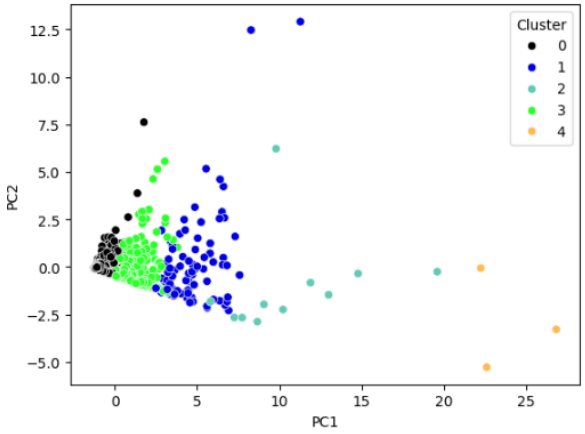
Gambar 13. Hasil reduksi data menggunakan PCA

Gambar 13 merupakan hasil reduksi dimensi data menggunakan metode PCA, data dibentuk menjadi 2 variabel atau 2 komponen utama, yaitu PC1 dan PC2.

	Attribute	PC_1	PC_2
0	Jumlah	0.314671	0.948017
1	Penjualan	0.482495	-0.170883
2	Penjualan Bersih	0.483846	-0.172573
3	HPP	0.481149	-0.176637
4	Laba Kotor	0.450077	-0.105263

Gambar 14. Hasil perhitungan vektor eigen pada metode PCA

Gambar 14 merupakan dari hasil vektor eigen, berdasarkan gambar diatas, atribut PC1 (komponen utama 1) terbentuk dari variabel/atribut Penjualan, Penjualan Bersih, HPP, dan Laba Kotor, menandakan bahwa variabel tersebut saling memiliki korelasi. Sedangkan atribut PC2 (komponen utama 2) terbentuk dari variable/atribut Jumlah saja.



Gambar 15. Visualisasi data hasil *clustering* menggunakan PCA

Gambar 15 merupakan visualisasi hasil *clustering k-means* menggunakan *scatter plot*.

```
result_size = result_df.groupby('Cluster').size()
print(result_size)
```

```
Cluster
0    1178
1      93
2      11
3     321
4       3
dtype: int64
```

Gambar 16. Jumlah data pada setiap *cluster*

Gambar 16 merupakan jumlah data pada setiap *cluster*, bisa dilihat bahwa Cluster 0 terdapat 1178 produk, Cluster 1 terdapat 93 produk, Cluster 2 terdapat 11 produk, Cluster 3 terdapat 321 produk, dan Cluster 4 terdapat 3 produk.

Setelah mengetahui persebaran produk pada setiap *cluster*, selanjutnya mengelompokan produk berdasarkan attribut *cluster*. Atribut Cluster didapat dari proses sebelumnya, yaitu pada proses pelabelan.

```
df_cluster0 = result_df[result_df['Cluster'] == 0]
df_cluster0

df_cluster1 = result_df[result_df['Cluster'] == 1]
df_cluster1

df_cluster2 = result_df[result_df['Cluster'] == 2]
df_cluster2

df_cluster3 = result_df[result_df['Cluster'] == 3]
df_cluster3

df_cluster4 = result_df[result_df['Cluster'] == 4]
df_cluster4
```

Gambar 17. Membuat *data frame* masing-masing *cluster*

Pada Gambar 17 produk sudah dipisah ke *data frame* masing-masing *cluster*, lalu dilakukan perbandingan rata-rata jumlah produk terjual dan rata-rata laba kotor dari produk yang terjual dari masing-masing *cluster*.

Tabel 4. Hasil pengelompokan produk Toko Agung Makmur Jaya

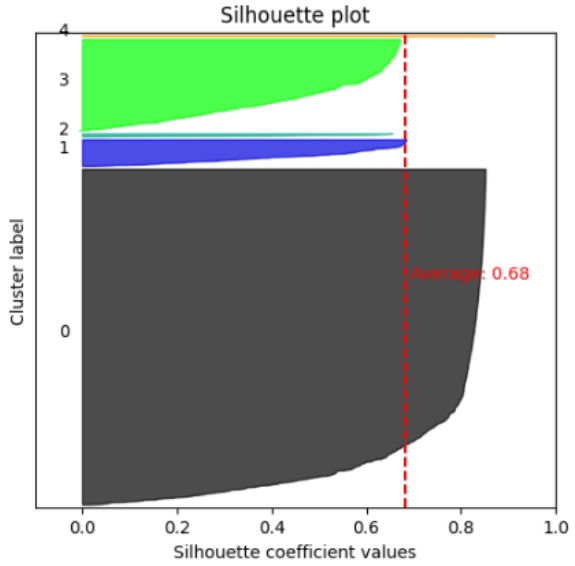
Cluster	**Nama Produk**	**Jumlah Varian Produk**	**Interpretasi *Cluster***
C0	Cetakan Twister Vip, Gelas Chs-9 Zf Bbc Mms, Lap Piring Murah Miki, Gm - 769 Gelas Minum, Piring Crystal Warna 6" Kh, …, Toples Susun 5 Iwatani Sm.	1178	Sangat Tidak Laku dan Omzet Sangat Tidak Besar
C1	Sendok Makan Water Mark Nk, Sendok Glowsy Nk, Sikat Eterna, Ts - 54 Bruno Sauce Keeper 425 Ml, Nh - 98 Minigo Bottle 600 Ml, ..., Panci Serbaguna 45 Cm Orchid.	93	Cukup Laku dan Omzet Cukup Besar
C2	Nampan Segi Stainless 555 No 27 Sm, Panci Fitri 18 Cm Jawa Maspion, Panci Larissa Maspion, M 030 Water Dispenser 10 Lt, K - 23 Water Jug 4,1 Ltr, …, Ap - 1 Thermos Panas " Air Pot " 2,5 Liter.	11	Laku dan Omzet Besar

Cluster	Nama Produk	Jumlah Varian Produk	Interpretasi *Cluster*
C3	Garpu Glowsy Nk, Irus Sop 218 Mj, Hanger Rotan Sweet, Lap Warna Arwana, T - 45 Bamboo Plate, …, I - 19 Marina Cooler Box 35 S (33 Ltr).	321	Tidak Laku dan Omzet Tidak Besar
C4	Sapu Ijuk Besar Dragon, H - 8 Siphon Pipe " Air Pot " 2,5 Liter, Sn - 8 Thermos Panas " Solaris " Sn 50 Hd.	3	Sangat Laku dan Omzet Sangat Besar

Tabel 4 merupakan interpretasi pada setiap *cluster*, intepretasi *cluster* didapat dari analisa setiap *cluster* berdasarkan variabel Jumlah dan Laba Kotor. Setiap *cluster* diurutkan berdasarkan perbandingkan rata-rata jumlah penjualan dan laba kotor. *Cluster* yang memiliki rata-rata penjualan dan rata-rata laba kotor nya paling tinggi dari *cluster* lain, maka *cluster* tersebut diinterpretasikan sebagai *cluster* dengan produk sangat laku dengan omzet sangat besar, begitu juga *cluster* lainnya dilakukan perbandingan, dari proses analisis tersebut *cluster-cluster* dapat diintepretasikan seperti pada tabel 4 diatas.

4.5. *Evaluation*

Fase *Evaluation* dilakukan untuk mengetahui seberapa baik hasil model *clustering k-means* yang telah dibuat. Hasil kualitas *cluster* yang dihasilkan dari algoritma *k-means* dievaluasi menggunakan *Silhoutte Coeffecient* dikatakan kuat apabila nilai indeksnya mendekati nilai 1 (satu).



Gambar 18. Grafik s*ilhoutte coeffecient* hasil c*lustering*

Gambar 18 merupakan grafik nilai *Silhoutte Coeffecient* dari setiap masing-masing *cluster* yang telah terbentuk pada fase *Modeling*. Adapun hasilnya berdasarkan grafik diatas, yaitu cluster 0 dan cluster 4 memiliki skor *Silhoutte* sekitar 0,8 yang artinya memiliki kualitas *cluster* kuat. Cluster 1, Cluster 2 dan Cluster 3 memiliki skor *Silhoutte* 0,6 yang artinya kulitas *cluster* tersebut standar. Secara keseluruhan hasil *clustering* dengan 5 *cluster* memiliki rata-rata skor *Silhoutte Coeffecient* sebesar 0,68 yang artinya keseluruhan hasil *clustering* k-*means* yang telah dilakukan memiliki struktur *cluster* dengan kualitas yang standar.

4.6. *Deployment*

Pada fase *Deployment* ini akan dilakukan pembuatan laporan mengenai pengetahuan yang didapat dari data penjualan yang telah diolah dari proses tahapan *data mining*. Hasil *deployment* dari penelitian ini, yaitu dalam bentuk laporan data analisis yang berisi informasi terkait pengelompokan produk. Dari informasi pengelompokan produk, pemilik Toko Agung Makmur Jaya bisa menggunakan pengetahuan tersebut sebagai pertimbangan strategi bisnis.

5. KESIMPULAN DAN SARAN

Kesimpulan yang didapat dari penelitian yang telah dilakukan adalah penerapan metode *clustering* dengan algoritma *k-means* untuk mengelompokan produk Toko Agung Makmur Jaya dilakukan dengan metodologi CRISP-DM (*Cross-Industry Standard Process for Data Mining*). Data yang digunakan adalah data penjualan produk bulan April–Juni 2024 (Kuartal II). Pengelompokan k-means terbentuk 5 *clutser* berdasarkan nilai DBI (*Davies-Bouldin Index*), yaitu sebesar 0,453559. Adapun hasil clustering, sebanyak 3 produk masuk ke dalam kelompok dengan penjualan sangat laku dan laba kotor sangat besar (C4), 11 produk masuk ke dalam kelompok dengan penjualan laku dan laba kotor besar (C2), 93 produk masuk ke dalam kelompok dengan penjualan cukup laku dan laba kotor cukup besar (C1), 321 produk masuk ke dalam kelompok dengan penjualan tidak laku dan laba kotor tidak besar (C3), dan 1178 produk masuk ke dalam kelompok dengan penjualan sangat tidak laku dan laba kotor sangat tidak besar (C0). Hasil *clustering* dievaluasi dengan *silhoutte coefficient*, didapat nilai 0,68. Berdasarkan kriteria nilai *silhoutte coeffecient*, *clustering* yang terbentuk memiliki struktur standar.

Saran untuk penelitian selanjutnya, dalam melakukan pemodelan *clustering* bisa menggunakan algoritma *k-medoids* atau algoritma yang lain. Diharapkan pada penelitian selanjutnya dalam mencari jumlah *cluster* yang optimal bisa menggunakan lebih dari satu metode agar lebih komprehensif.

DAFTAR PUSTAKA

[1] Syahril, M., Erwansyah, K. and Yetri, M., "Penerapan Data Mining Untuk Menentukan Pola Penjualan Peralatan Sekolah Pada Brand Wigglo Dengan Menggunakan Algortima Apriori," *J-SISKO TECH* (*Jurnal Teknologi Sistem Informasi Dan Sistem Komputer TGD),*

vol. 3, no. 1, pp. 118–136, Jan. 2020, doi: 10.53513/jsk.v3i1.202.

[2] Astuti, N., Utamajaya, J. N. and Pratama, A., "Penerapan Data Mining Pada Penjualan Produk Digital Konter Leppangeng Cell Menggunakan Metode K-Means Clustering," *JURIKOM (Jurnal Riset Komputer)*, vol. 9, no. 3, pp. 754–760, Jun. 2022, doi: 10.30865/jurikom.v9i3.4351.

[3] Rismayadi, A. A., Fatonah, N. N. and Junianto, E., "Algoritma K-Means Clustering Untuk Menentukan Strategi Pemasaran di Cv. Integreet Konstruksi," *JURNAL RESPONSIF*, vol. 3, no. 1, pp. 30–36, 2021.

[4] Aditya, A., Jovian, I. and Sari, B. N., "Implementasi K-Means Clustering Ujian Nasional Sekolah Menengah Pertama di Indonesia Tahun 2018/2019," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 4, no. 1, pp. 51–58, Jan. 2020, doi: 10.30865/mib.v4i1.1784.

[5] Pangestu, B., Kristiawan, N. and Sulistiyowati, N., "Clustering Obat Untuk Menentukan Pola Pemasaran Efektif di Apotek Amarta Sehat," *Jurnal Wahana Pendidikan*, vol. 8, no. 16, pp. 115–126, Sep. 2022.

[6] Muningsih, E., Maryani, I. and Handayani, V. R., "Penerapan Metode K-Means dan Optimasi Jumlah Cluster dengan Index Davies Bouldin untuk Clustering Propinsi Berdasarkan Potensi Desa," *Evolusi: Jurnal Sains Dan Manajemen*, vol. 9, no. 1, pp. 95–100, Mar. 2021, doi: 10.31294/evolusi.v9i1.10428.

[7] Rahmawati, E., Sari, B. N. and Jajuli, M., "Implementasi Algoritma K-Means pada Pemetaan Daerah Terdampak Tanah Longsor di Jawa Tengah," *JATI (Junal Mahasiswa Teknik Informatika)*, vol. 8, no. 3, pp. 2568–2574, Jun. 2024, doi: 10.36040/jati.v8i3.9487.

[8] Arhami, M. and Nasir, M., *Data Mining: Algoritma dan Iplementasi*. Yogyakarta: ANDI, 2020.

[9] Vulandari, R. T., *Data Mining: Teori dan Aplikasi Rapidminer*. Yogyakarta: GAVA MEDIA, 2017.

[10] Winarta, A. and Kurniawan, W. J., "Optimasi Cluster K-Means Menggunakan Metode Elbow Pada Data Pengguna Narkoba Dengan Pemrograman Python," *(JTIK) Jurnal Teknik Informatika Kaputama*, vol. 5, no. 1, pp. 113–119, Jan. 2021, doi: 10.59697/jtik.v5i1.593.

[11] Butsianto, S. and Saepudin, N., "Penerapan Data Mining Terhadap Minat Siswa Dalam Mata Pelajaran Matematika Dengan Metode K-Means," *Jurnal Nasional Komputasi Dan Teknologi Informasi (JNKTI)*, vol. 3, no. 1, pp. 51–59, 2020, doi: 10.32672/jnkti.v3i1.2008.

[12] Dewi, D. A. I. C. and Pramita, D. A. K., "Analisis Perbandingan Metode Elbow dan Sillhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali," *Matrix: Jurnal Manajemen Teknologi dan Informatika*, vol. 9, no. 3, 2019, doi: 10.31940/matrix.v9i3.1662.

[13] Farissa, R. A., Mayasari, R. and Umaidah, Y., "Perbandingan Algoritma K-Means dan K-Medoids Untuk Pengelompokkan Data Obat dengan Silhouette Coefficient di Puskesmas Karangsambung," *Journal of Applied Informatics and Computing (JAIC)*, vol. 5, no. 2, pp. 109–116, Dec. 2021, doi: 10.30871/jaic.v5i1.3237.

[14] Manalu, D. A. and Gunadi, G., "Implementasi Metode Data Mining K-Means Clustering Terhadap Data Pembayaran Transaksi Menggunakan Bahasa Pemrograman Python Pada Cv Digital Dimensi," *Infotech: Journal of Technology Information*, vol. 8, no. 1, pp. 43–54, 2022, doi: 10.37365/jti.v8i1.131.

[15] Soen, G. I. E., Marlina and Renny, "Implementasi Cloud Computing dengan Google Colaboratory Pada Aplikasi Pengolah Data Zoom Participants," *JITU: Journal Informatic Technology And Communication*, vol. 6, no. 1, 2022.

[16] Kristiawan, N. A., "Penerapan Algoritma K-Means Clustering Untuk Analisis Dan Pemetaan Daerah Rawan Stunting di Kabupaten Karawang," B.S. Thesis, Faculty of Comp. Science., Singaperbangsa Univ., Karawang, 2022.