

**OPTIMASI KLASTERING K-MEANS UNTUK SEGMENTASI
PELANGGAN DENGAN INTEGRASI PCA, DBSCAN DAN K-
MEANS++**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi S1 Informatika



disusun oleh

Tonik Purwanto

20.11.3652

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2026

**OPTIMASI KLASTERING K-MEANS UNTUK SEGMENTASI
PELANGGAN DENGAN INTEGRASI PCA, DBSCAN DAN K-
MEANS++**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi S1 Informatika



disusun oleh
Tonik Purwanto
20.11.3652

Kepada

FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026

HALAMAN PERSETUJUAN

SKRIPSI

**OPTIMASI KLASTERING K-MEANS UNTUK SEGMENTASI
PELANGGAN DENGAN INTEGRASI PCA, DBSCAN DAN K-MEANS++**

yang disusun dan diajukan oleh

Tonik Purwanto

20.11.3652

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 09 Mei 2026

Dosen Pembimbing,



Windha Mega Pradnya Dhuhita, S.Kom., M.Kom.
NIK. 190302185

HALAMAN PENGESAHAN

SKRIPSI

**OPTIMASI KLASTERING K-MEANS UNTUK SEGMENTASI
PELANGGAN DENGAN INTEGRASI PCA, DBSCAN DAN K-MEANS++**

yang disusun dan diajukan oleh

Tonik Purwanto

20.11.3652

Telah dipertahankan di depan Dewan Penguji
pada tanggal 21 Mei 2026

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Theopilus Bayu Sasongko, S.Kom., M.Eng.
NIK. 190302375

Afrig Aminuddin, S.Kom., M.Eng., Ph.D.
NIK. 190302351

Windha Mega Pradnya Dhuhita, S.Kom., M.Kom.
NIK. 190302185

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 21 Mei 2026

DEKAN FAKULTAS ILMU KOMPUTER



Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Tonik Purwanto
NIM : 20.11.3652

Menyatakan bahwa Skripsi dengan judul berikut:

Optimasi Klustering K-Means untuk Segmentasi Pelanggan dengan Integrasi PCA, DBSCAN dan K-Means++

Dosen Pembimbing : Windha Mega Pradnya Duhita, S.Kom., M.Kom.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 21 Mei 2026

Yang Menyatakan,

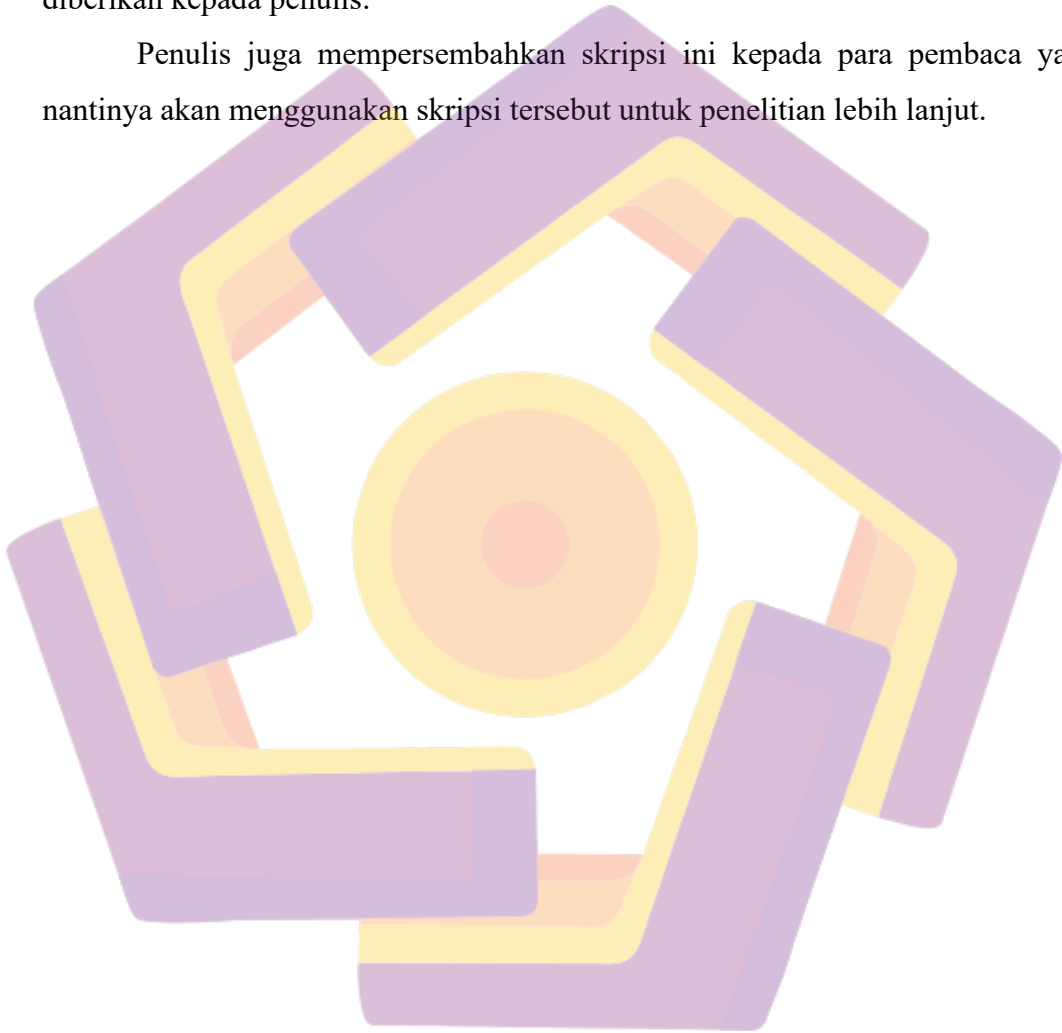


Tonik Purwanto

HALAMAN PERSEMBAHAN

Dengan mengucapkan puji syukur kepada Allah SWT, penulis mempersembahkan karya ini kepada kedua orang tua dan keluarga tercinta yang telah memberikan dukungan, kasih sayang dan untaian doa yang tiada henti yang diberikan kepada penulis.

Penulis juga mempersembahkan skripsi ini kepada para pembaca yang nantinya akan menggunakan skripsi tersebut untuk penelitian lebih lanjut.



KATA PENGANTAR

Assalamualaikum Warahmatullahi Wabarakatuh

Puji dan syukur atas kehadiran Allah SWT yang telah melimpahkan rahmat, taufik serta hidayah-Nya sehingga penulis dapat menyelesaikan skripsi ini dengan baik. Shalawat serta salam selalu tercurahkan kepada junjungan Nabi Muhammad SAW yang telah membawa umat manusia keluar dari zaman kegelapan.

Dalam penulisan skripsi ini, penulis menyadari bahwa masih banyak terdapat kekurangan. Terlepas dari hal tersebut, skripsi ini dapat memenuhi salah satu syarat untuk memperoleh gelar Sarjana Komputer strata satu (S1) Program Studi Informatika di Fakultas Ilmu Komputer Universitas Amikom Yogyakarta. Skripsi yang berjudul “Optimasi Klastering K-Means untuk Segmentasi Pelanggan dengan Integrasi PCA, DBSCAN dan K-Means++” akhirnya dapat diselesaikan dengan baik sesuai harapan penulis. Selama penyusunan skripsi ini, penulis menghadapi berbagai hambatan dan tantangan.

Pada kesempatan ini, penulis ingin menyampaikan rasa terimakasih kepada seluruh pihak yang telah memberikan dukungan, sehingga penulis dapat menyelesaikan skripsi ini. Secara khusus, penulis menyampaikan ucapan terimakasih kepada:

1. Bapak Prof. Dr. Mohammad Suyanto, M.M. selaku Rektor Universitas Amikom Yogyakarta.
2. Ibu Prof. Dr. Kusrini, S.Kom., M.Kom. selaku Dekan Fakultas Ilmu Komputer Universitas Amikom Yogyakarta.
3. Ibu Eli Pujastuti, M.Kom. selaku Ketua Program Studi Informatika Universitas Amikom Yogyakarta.
4. Bapak Arif Akbarul Huda, S.Si., M.Eng. selaku Sekretaris Program Studi Informatika Universitas Amikom Yogyakarta.
5. Ibu Windha Mega Pradnya Duhita, S.Kom., M.Kom. selaku Dosen Pembimbing yang telah menyediakan waktunya untuk memberikan ilmu dan tentunya dukungan kepada penulis.

6. Dosen-dosen Program Studi Informatika yang telah memberikan ilmu selama perkuliahan.
7. Seluruh staf Fakultas Ilmu Komputer yang telah membantu penulis selama perkuliahan.
8. Kedua orang tua dan keluarga tercinta yang selalu memberikan doa, dukungan, semangat dan kasih sayang kepada penulis.
9. Teman-teman dan semua pihak yang tidak dapat disebut satu per satu yang telah menemani peneliti dari awal perkuliahan hingga laporan skripsi ini terselesaikan.

Penulis menyadari bahwa dalam penyusunan skripsi ini masih jauh dari kata sempurna. Oleh karena itu, penulis mengharapkan kritik dan saran yang membangun demi penyempurnaan karya ini dimasa mendatang. Akhir kata, semoga skripsi ini dapat memberikan manfaat bagi pembaca dan penulis sendiri.

Wassalamualaikum Warahmatullahi Wabarakatuh.

Yogyakarta, 21 Mei 2026

Penulis

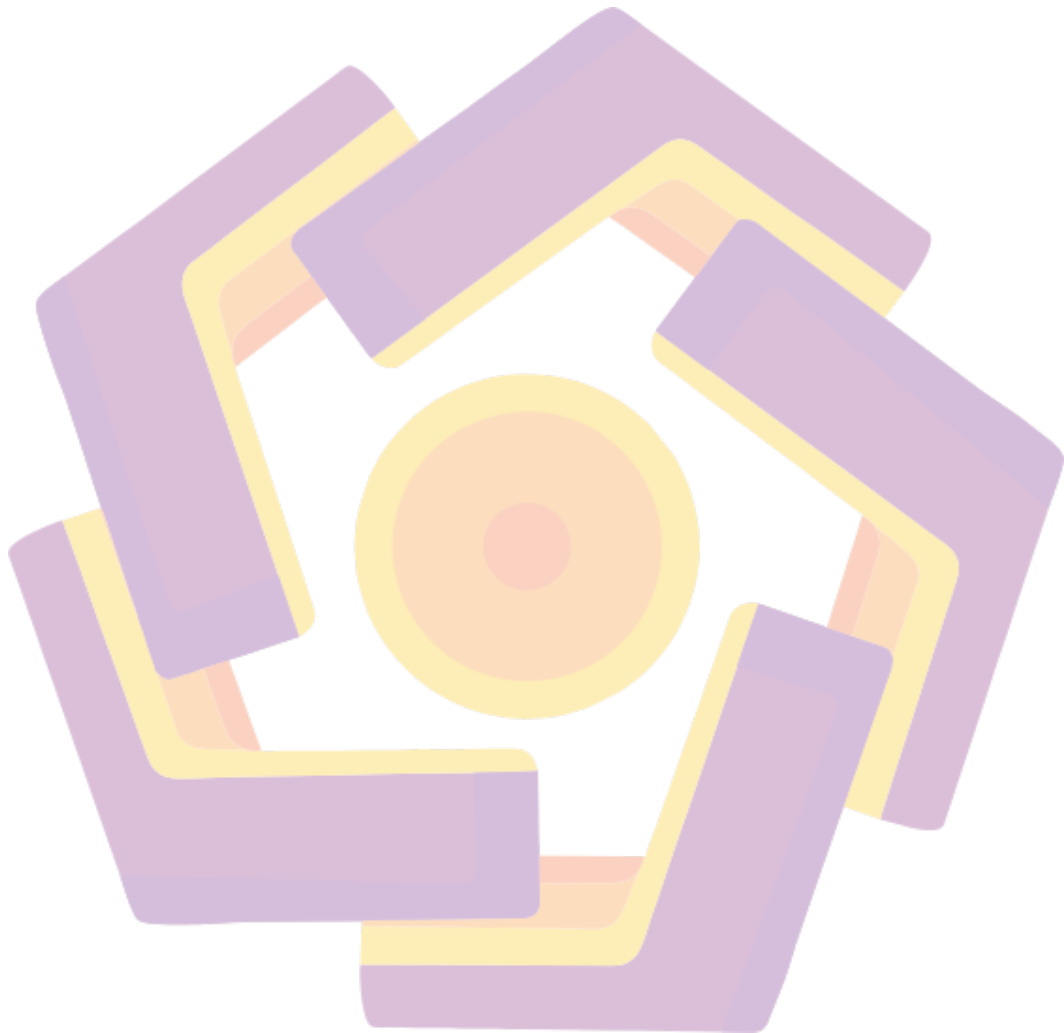
DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PERSETUJUAN.....	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI	iv
HALAMAN PERSEMBAHAN	v
KATA PENGANTAR	vi
DAFTAR ISI.....	viii
DAFTAR TABEL.....	xii
DAFTAR GAMBAR.....	xiii
DAFTAR LAMPIRAN.....	xiv
DAFTAR LAMBANG DAN SINGKATAN	xv
DAFTAR ISTILAH.....	xviii
INTISARI	xxi
<i>ABSTRACT</i>	xxii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	4
1.3 Batasan Masalah	4
1.4 Tujuan Penelitian	4
1.5 Manfaat Penelitian	5
1.5.1 Manfaat Teoritis.....	5
1.5.2 Manfaat Praktis	5
1.6 Sistematika Penulisan	5

BAB II TINJAUAN PUSTAKA	7
2.1 Studi Literatur	7
2.2 Dasar Teori.....	18
2.2.1 Data Mining	18
2.2.2 Segmentasi Pelanggan	18
2.2.3 RFM	18
2.2.4 Klasterisasi.....	21
2.2.5 Algoritma K-Means	21
2.2.6 Algoritma K-Means++.....	23
2.2.7 Principal Component Analysis	24
2.2.8 DBSCAN	24
2.2.9 IQR.....	25
2.2.10 Metrik Evaluasi.....	26
2.2.11 CRISP-DM.....	29
BAB III METODE PENELITIAN	33
3.1 Objek Penelitian.....	33
3.2 Alur Penelitian	33
3.2.1 Business Understanding	34
3.2.2 Data Understanding	35
3.2.3 Data Preparation.....	37
3.2.4 Modeling	38
3.2.5 Evaluation	39
3.3 Alat dan Bahan.....	40
3.3.1 Data Penelitian	40
3.3.2 Perangkat Lunak	40

3.3.3	Perangkat Keras	40
BAB IV HASIL DAN PEMBAHASAN		41
4.1	<i>Business Understanding</i>	41
4.2	<i>Data Understanding</i>	41
4.2.1	Pengumpulan data	41
4.2.2	Deskripsi Dataset	42
4.2.3	Eksplorasi Data	47
4.2.4	Hasil Identifikasi Permasalahan Data	53
4.2.5	Keterkaitan RFM	53
4.3	<i>Data Preparation</i>	55
4.3.1	Data Selection	55
4.3.2	Data Cleaning	56
4.3.3	Feature Engineering	57
4.3.4	Data Transformation	61
4.4	Modeling	64
4.4.1	Modeling K-Means	64
4.4.2	Modeling K-Means++	71
4.4.3	Modeling Hybrid	81
4.5	Evaluation	92
4.5.1	Tujuan Evaluasi	92
4.5.2	Validasi Jumlah Kluster Optimal	92
4.5.3	Perbandingan Model Klasterisasi	93
4.5.4	Penentuan Model Terbaik	95
4.5.5	Interpretasi Model Terbaik	95
BAB V PENUTUP		97

5.1	Kesimpulan	97
5.2	Saran	98
REFERENSI		99
LAMPIRAN.....		106



DAFTAR TABEL

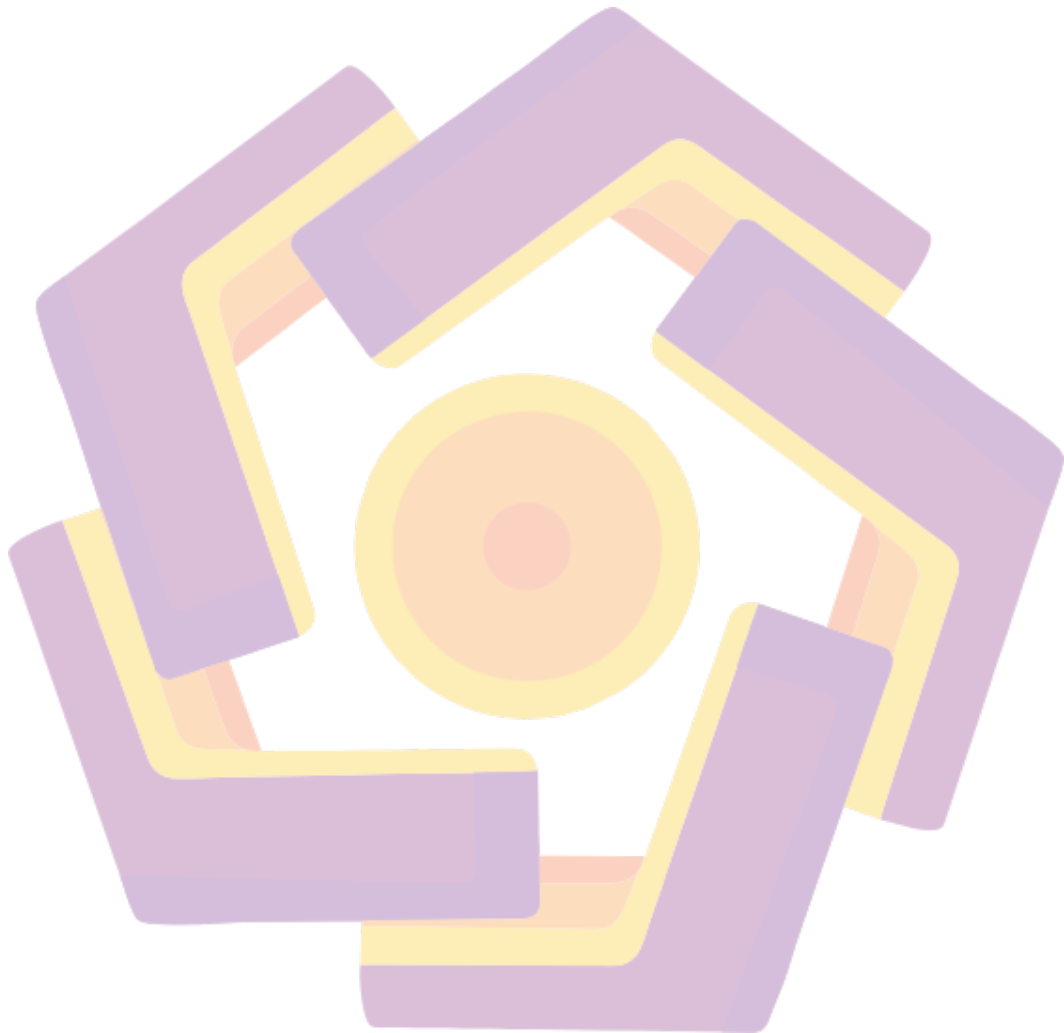
Tabel 2. 1 Keaslian Penelitian	10
Tabel 4. 1 Informasi Dataset	43
Tabel 4. 2 Hasil Statistik Deskriptif.....	44
Tabel 4. 3 Tampilan Data.....	46
Tabel 4. 4 Lanjutan Tampilan Data	46
Tabel 4. 5 Missing Value	48
Tabel 4. 6 Data Duplikat	49
Tabel 4. 7 Data Tidak Valid.....	49
Tabel 4. 8 Data Retur	50
Tabel 4. 9 Outlier	52
Tabel 4. 10 Keterkaitan RFM	54
Tabel 4. 11 Data Selection	56
Tabel 4. 12 Data Cleaning	56
Tabel 4. 13 Pembentukan TotalPrice	57
Tabel 4. 14 Pembentukan RFM	59
Tabel 4. 15 Transformasi Logaritmik	61
Tabel 4. 16 Perhitungan Mean dan Standar Deviasi.....	62
Tabel 4. 17 Normalisasi	63
Tabel 4. 18 Pengujian Jumlah Klaster K-Means	65
Tabel 4. 19 Pembentukan Model K-Means	70
Tabel 4. 20 Pengujian Jumlah Klaster K-Means++	72
Tabel 4. 21 Pembentukan Model K-Means++	77
Tabel 4. 22 Deteksi Outlier	81
Tabel 4. 23 Perhitungan Jumlah Outlier	83
Tabel 4. 24 Pengujian Jumlah Klaster Hybrid	85
Tabel 4. 25 Pembentukan Model K-Means++ Hybrid.....	90
Tabel 4. 26 Analisis Kualitas Klasterisasi	93
Tabel 4. 27 Interpretasi Model Terbaik	95

DAFTAR GAMBAR

Gambar 2. 1 Alur Metode CRISP-DM	30
Gambar 3. 1 Alur Penelitian	35
Gambar 4. 1 Proses Pengumpulan Data	41
Gambar 4. 2 Eksplorasi Data	47
Gambar 4. 3 Distribusi Quantity	51
Gambar 4. 4 Distribusi UnitPrice	52
Gambar 4. 5 Data Preparation	55
Gambar 4. 6 Analisis Korelasi RFM	60
Gambar 4. 7 Silhouette Score K-Means	66
Gambar 4. 8 Davies-Bouldin Index K-Means	67
Gambar 4. 9 Calinski-Harabasz Index K-Means	70
Gambar 4. 10 Klastering K-Means	71
Gambar 4. 11 Silhouette Score K-Means++	74
Gambar 4. 12 Davies-Bouldin Index K-Means++	75
Gambar 4. 13 Calinski-Harabasz Index K-Means++	76
Gambar 4. 14 Klastering K-Means++	80
Gambar 4. 15 Hasil PCA	80
Gambar 4. 16 Deteksi Outlier	82
Gambar 4. 17 Penghapusan Outlier	84
Gambar 4. 18 Silhouette Score Hybrid	87
Gambar 4. 19 Davies-Bouldin Index Hybrid	90
Gambar 4. 20 Calinski-Harabasz Index Hybrid	91
Gambar 4. 21 Klastering Hybrid	91
Gambar 4. 22 Perbandingan Model Klasterisasi	94

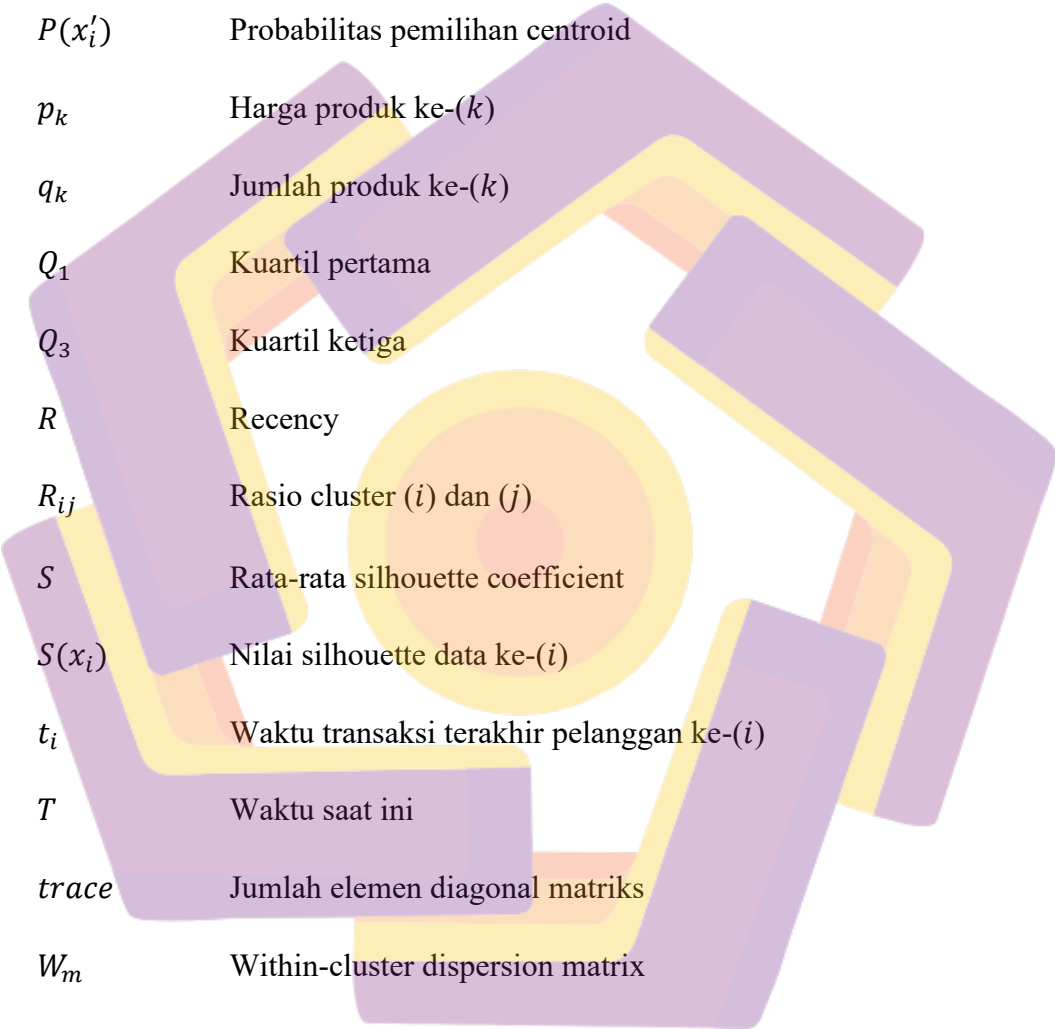
DAFTAR LAMPIRAN

Lampiran. 1 Surat Pembimbing Skripsi.....	106
Lampiran. 2 Source Code Google Colaboratory.....	108



DAFTAR LAMBANG DAN SINGKATAN

$(.)^T$	Transpose matriks
a_i	Rata-rata jarak intra-cluster
A	Himpunan centroid
avg_i	Rata-rata jarak cluster (i)
avg_j	Rata-rata jarak cluster (j)
b_i	Rata-rata jarak inter-cluster terdekat
B_m	Between-cluster dispersion matrix
C	Jumlah cluster
c_i	Centroid cluster ke-(i)
$CH(c)$	Nilai Calinski-Harabasz Index
$D(x_i)$	Jarak minimum data ke centroid
$d(x, y)$	Jarak antara data (x) dan (y)
d_{ij}	Jarak centroid cluster (i) dan (j)
$dist(x, y)$	Fungsi jarak antara dua data
F	Frequency
IQR	Interquartile Range
k	Indeks cluster
K	Jumlah cluster
M	Monetary
$MinObj$	Fungsi objektif minimum
μ	Rata-rata seluruh data

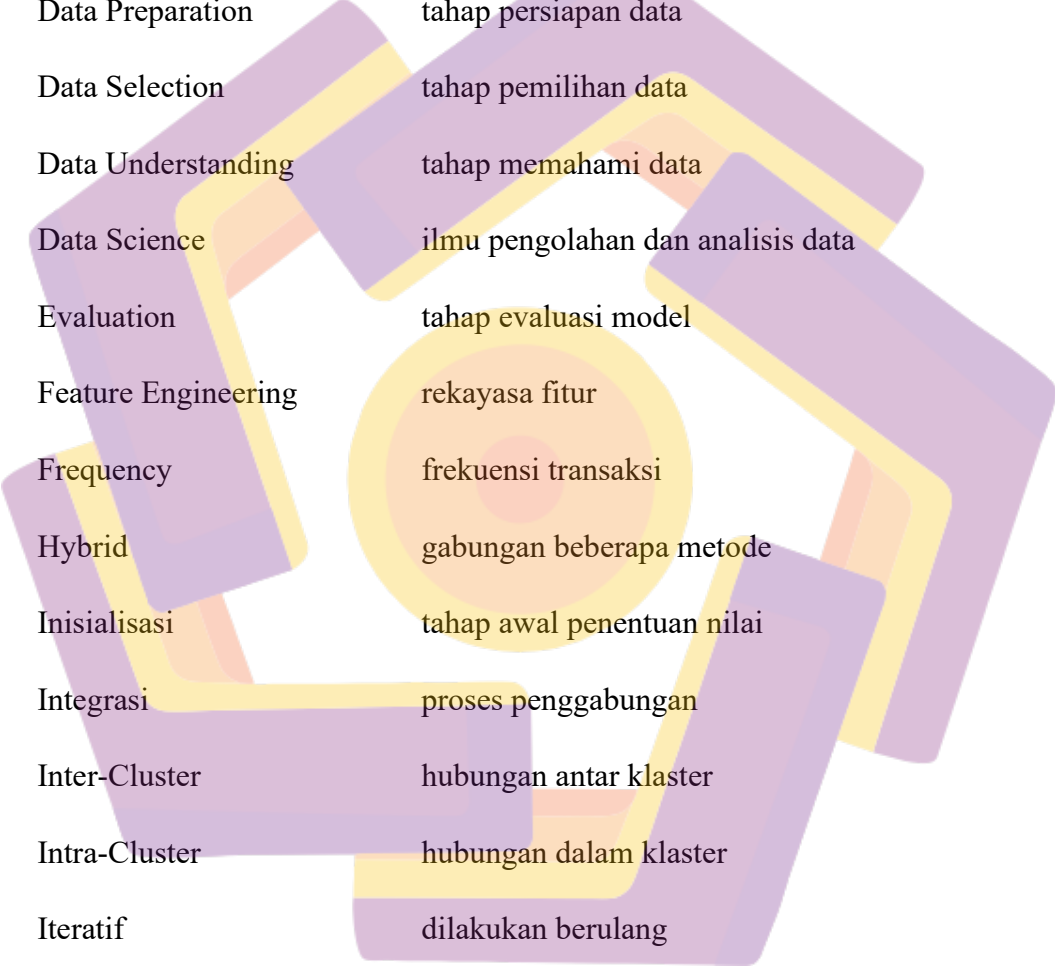


n	Jumlah transaksi
N	Jumlah data
n_i	Jumlah anggota cluster ke- (i)
$OptK$	Jumlah cluster optimal
$P(x'_i)$	Probabilitas pemilihan centroid
p_k	Harga produk ke- (k)
q_k	Jumlah produk ke- (k)
Q_1	Kuartil pertama
Q_3	Kuartil ketiga
R	Recency
R_{ij}	Rasio cluster (i) dan (j)
S	Rata-rata silhouette coefficient
$S(x_i)$	Nilai silhouette data ke- (i)
t_i	Waktu transaksi terakhir pelanggan ke- (i)
T	Waktu saat ini
$trace$	Jumlah elemen diagonal matriks
W_m	Within-cluster dispersion matrix
x_i	Data ke- (i)
X	Himpunan data
y_i	Data pembanding ke- (i)
ε	Radius tetangga pada DBSCAN
CH Index	Calinski-Harabasz Index

CRISP-DM	Cross Industry Standard Process for Data Mining
DBI	Davies-Bouldin Index
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
IQR	Interquartile Range
KDD	Knowledge Discovery in Databases
PCA	Principal Component Analysis
RFM	Recency, Frequency, Monetary



DAFTAR ISTILAH



Baseline	nilai atau model dasar pembanding
Business Understanding	tahap memahami kebutuhan bisnis
Centroid	titik pusat klaster
Data Mining	proses penggalian informasi dari data
Data Preparation	tahap persiapan data
Data Selection	tahap pemilihan data
Data Understanding	tahap memahami data
Data Science	ilmu pengolahan dan analisis data
Evaluation	tahap evaluasi model
Feature Engineering	rekayasa fitur
Frequency	frekuensi transaksi
Hybrid	gabungan beberapa metode
Inisialisasi	tahap awal penentuan nilai
Integrasi	proses penggabungan
Inter-Cluster	hubungan antar klaster
Intra-Cluster	hubungan dalam klaster
Iteratif	dilakukan berulang
Kontinu	data berkelanjutan
konvergensi	kondisi stabil hasil iterasi
Kovarians	ukuran hubungan antar variabel
Kuartil	pembagian data menjadi empat bagian
Linier	berbentuk garis lurus

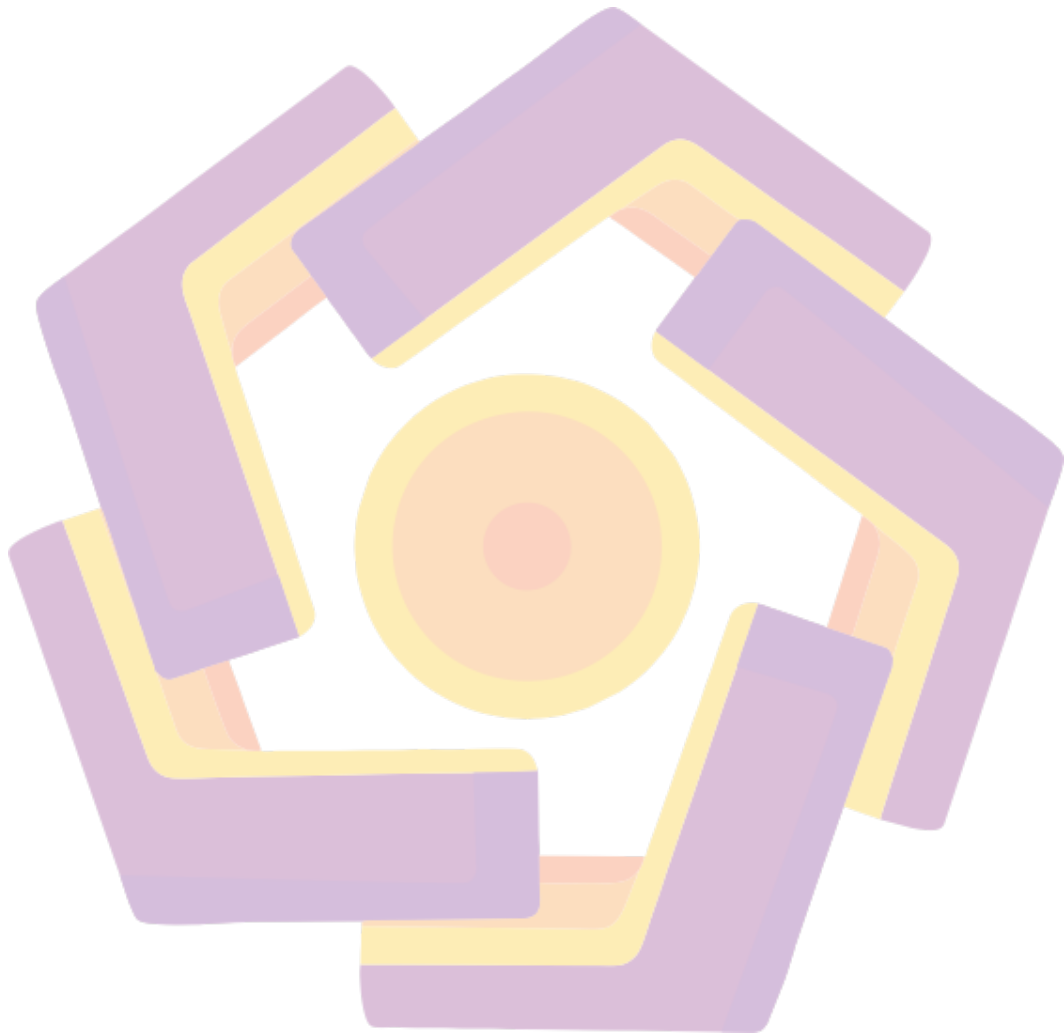
Machine Learning	pembelajaran mesin
Metrik	ukuran evaluasi
Missing Value	data yang hilang
Modeling	tahap pembangunan model
Monetary	nilai total transaksi
Multivariat	analisis banyak variabel
Noise	gangguan atau data tidak relevan
Non-Parametrik	metode tanpa asumsi distribusi tertentu
Objek	entitas yang dianalisis
Ortogonal	saling tegak lurus
Outlier	data pencilan
Probabilitas	peluang kejadian
Real-World	kondisi dunia nyata
Recency	kedekatan waktu transaksi terakhir
Robust	tahan terhadap gangguan atau <i>outlier</i>
Similarity	tingkat kemiripan
Skewness	tingkat kemencengan distribusi data
Standarisasi	proses penyetaraan skala data
Transformasi Logaritmik	transformasi menggunakan logaritma
Transformasi	perubahan bentuk data
Unsupervised Learning	pembelajaran tanpa label
Univariat	analisis satu variabel

Variabel

tempat menyimpan data

Vektor

besaran yang memiliki arah dan nilai



INTISARI

Segmentasi pelanggan merupakan salah satu strategi penting dalam memahami perilaku pelanggan dan mendukung pengambilan keputusan bisnis. Namun, metode klasterisasi konvensional seperti K-Means memiliki kelemahan dalam menentukan jumlah klaster secara optimal, sensitif terhadap inisialisasi titik pusat (*centroid*) awal, serta kurang mampu menangani penciran (*outlier*) pada data. Permasalahan tersebut dapat menyebabkan kualitas segmentasi pelanggan menjadi kurang optimal, sehingga strategi pemasaran, pelayanan pelanggan dan pengambilan keputusan bisnis menjadi kurang tepat sasaran.

Penelitian ini menggunakan metode eksperimen dengan pendekatan CRISP-DM untuk menyelesaikan permasalahan tersebut. Tahapan penelitian dimulai dari *business understanding*, *data understanding*, *data preparation*, *modeling* dan *evaluation*. *dataset* yang digunakan adalah Online Retail dengan pendekatan RFM (*recency*, *frequency*, *monetary*). Metode yang diusulkan adalah *hybrid* dengan mengintegrasikan PCA untuk menyederhanakan struktur data, mengurangi korelasi antar fitur, serta membantu visualisasi data klasterisasi, DBSCAN untuk mengidentifikasi *noise* yang merepresentasikan *outlier* dan menentukan jumlah klaster secara otomatis, serta K-Means++ untuk menghasilkan klaster akhir yang lebih stabil.

Hasil penelitian menunjukkan bahwa metode *hybrid* menghasilkan performa klasterisasi terbaik dengan nilai Silhouette Score sebesar 0.464281, lebih tinggi dibandingkan K-Means dan K-Means++ sebesar 0.432940. Selain itu, metode *hybrid* menghasilkan Davies-Bouldin Index (DBI) terendah sebesar 0.821916 dan Calinski-Harabasz Index tertinggi sebesar 5103.759768. Kontribusi penelitian ini adalah memberikan metode segmentasi pelanggan yang lebih optimal untuk membantu perusahaan ritel dalam menentukan strategi pemasaran dan pengelolaan pelanggan.

Kata kunci: Segmentasi Pelanggan, Klasterisasi, K-Means++, PCA, DBSCAN

ABSTRACT

Customer segmentation is an important strategy for understanding customer behavior and supporting business decision-making. However, conventional clustering methods such as K-Means have several limitations, including difficulty in determining the optimal number of clusters, sensitivity to the initial centroid initialization, and limited capability in handling outliers in the dataset. These issues may lead to suboptimal customer segmentation results, thereby reducing the effectiveness of marketing strategies, customer service, and business decision-making.

This study adopts an experimental approach using the CRISP-DM methodology to address these challenges. The research process consists of business understanding, data understanding, data preparation, modeling, and evaluation stages. The dataset used in this study is the Online Retail dataset with the RFM (Recency, Frequency, Monetary) approach. The proposed method is a hybrid model that integrates PCA to simplify data structure, reduce inter-feature correlation, and support clustering visualization; DBSCAN to identify noise representing outliers and automatically estimate the number of clusters; and K-Means++ to produce more stable final clusters.

The results show that the proposed hybrid method achieves the best clustering performance with a Silhouette Score of 0.464281, which is higher than K-Means and K-Means++ at 0.432940. In addition, the hybrid method produces the lowest Davies-Bouldin Index (0.821916) and the highest Calinski-Harabasz Index (5103.759768). The contribution of this research is the development of a more optimal customer segmentation method to assist retail companies in determining marketing strategies and customer management.

Keyword: Customer Segmentation, Clustering, K-Means++, PCA, DBSCAN