# Penerapan *Machine Learning* untuk Pencarian Pelanggan Loyal Berpotensi Menggunakan Metode Python Pandas Seaborn

## *Application of Machine Learning for Identifying Potential Loyal Customers Using Python Pandas Seaborn Method*

Yustina Tritularsih[1)*], Hoedi Prasetyo[2)]
[1)] Prodi Teknik Mekatronika, Politeknik ATMI, Surakarta, Indonesia
[2)] Prodi Teknologi Rekayasa Mekatronika, Politeknik ATMI, Surakarta, Indonesia
email: [1)*]yustina_tritularsih@atmi.ac.id, [2)]hoedi.prasetyo@atmi.ac.id

**Informasi Artikel**
Diterima:
*Submitted:*
25/09/2024

Diperbaiki:
*Revised:*
24/01/2025

Disetujui:
*Accepted:*
30/01/2025

[*)] yustina tritularsih
yustina_tritularsih@atmi.ac.id

DOI:
https://doi.org/10.32502/intengrasi.v10i1.292

**Abstrak**

Kunci kesuksesan berbisnis agar target penjualan tercapai dengan menjaga kepercayaan dan minat daya beli konsumen untuk tetap menjadi pelanggan setia. Salah satu faktor penunjangnya di bidang pemasaran yaitu dengan cara memprediksi untuk pencarian pelanggan loyal berpotensi. Untuk membangun model tersebut, perlu data demografi, sosial, transaksional, metrik perilaku, dan fitur pendukung lainnya. Masalah utama yang terjadi saat ini terbatasnya bagian pemasaran memiliki data pelanggan dan hanya mengandalkan informasi yang disediakan oleh sistem ERP yang sebagian besar datanya berorientasi transaksional. Tujuan utama dari penelitian ini adalah untuk mengusulkan kombinasi analisis RFM *(Recency, Frequency and Monetary)* dan algoritma *machine learning* untuk memprediksi potensi pelanggan berdasarkan sebagian besar data transaksional. Dataset diambil dari sistem ERP perusahaan sheet metal di PT.ABC. Skor RFM dihitung untuk setiap pelanggan dalam jangka waktu 6 bulan sebelum tanggal akhir pemeriksaan. Nilai target untuk model prediksi ini adalah metrik pelanggan berpotensi yang menunjukkan apakah pelanggan telah melakukan transaksi dalam 6 bulan ke depan setelah analisis RFM atau tidak. Eksperimen dilakukan dengan metode Python Pandas Seaborn. Hasil menunjukkan batasan skor dan metrik RFM menggunakan algoritma *machine learning*, perusahaan dapat memprediksi pelanggan loyal berpotensi. Skor terbaik ditunjukkan pelanggan platinum cenderung berbelanja lebih sering dan lebih banyak berbelanja dibandingkan pelanggan lain.

**Kata kunci**: Prediksi, RFM, Pelanggan Loyal, Transaksi, Algoritma.

***Abstract***

*The key to business success in achieving sales targets is to maintain consumer trust and purchasing interest to keep them as loyal customers. One of the supporting factors in the field of marketing is predicting potential loyal customers. To build that model, demographic, social, transactional data, behavioural metrics, and other supporting features are needed. The main issue currently is that the marketing department has limited customer data and relies solely on the information provided by the ERP system, which is mostly transaction-oriented. The main objective of this research is to propose a combination of RFM (Recency, Frequency, and monetary) analysis and machine learning algorithms to predict customer potential based on a majority of transactional data. The dataset was taken from the ERP system of the sheet metal company PT.ABC. The RFM score is calculated for each customer over a period of 6 months prior to the end date of the assessment. The target value for this predictive model is a customer metric that indicates whether the customer has made a transaction within 6 months following the RFM analysis or not. The experiment was conducted using Python Pandas Seaborn methods. The results show that by using machine learning algorithms to establish score limits and RFM metrics, companies can predict*

*potential loyal customers. Platinum customers tend to score the best as they shop more frequently and spend more compared to other customers.*
.

***Keywords**: Predicting, RFM, Loyal Customers, Transactional, Algorithms.*


p-ISSN *2528-7419*
e-ISSN *2654-5551*

## Pendahuluan

Analisis RFM *(Recency, Frequency and Monetary)* adalah pendekatan klasik untuk penilaian dan segmentasi pelanggan. Metode ini berhasil diterapkan selama beberapa dekade yang menerapkan konsep segmentasi pelanggan berdasarkan frekuensi jumlah pembelian yang lebih banyak, lebih sering melakukan transaksi baru, atau yang akan segera melakukan transaksi dalam waktu dekat. Adanya kemajuan teknologi kian pesat saat ini, banyak peneliti mengolah data pelanggan menggunakan aplikasi *machine learning* untuk mengimplementasikan lebih cepat dalam model pencarian data yang diharapkan [2]. Selain itu dengan *machine learning* data yang dihasilkan berkualitas tinggi dan dapat diklasifikasikan untuk mendapatkan prediksi seperti pelanggan setia, pelanggan berpotensi dan nilai pendapatan yang diperoleh [7], [10].

Menurut Ahmed dan Patil Saurabh, mengevaluasi dengan menggunakan klasifikasi Naïve Bayes dan Neural Network sebagai metode klasifikasi untuk memprediksi dan mempertahankan pelanggan yang setia adalah paling menguntungkan [1]. Analisis RFM sering direkomendasikan untuk dilakukan dengan metode pengelompokan dengan menggunakan algoritma klastering *(K-means)* untuk mengeksplorasi pengelompokan dari pelanggan berpotensi dan pelanggan setia [18]. Penelitian lain yang dilakukan Dewabharata, menemukan dengan metode pohon keputusan, jaringan neural, dan regresi linier dapat digunakan sebagai metode prediksi pencarian pelanggan berpotensi melalui metode yang telah mengusulkan kerangka kerja untuk menggabungkan analisis RFM dengan metode dan pola urutan untuk mengidentifikasi klien VIP dan memprediksi pendapatan yang akan datang [6]. Model yang paling sederhana lainnya yang telah banyak dilakukan peneliti untuk mengatasi kelemahan analisis RFM, seperti hanya berfokus pada perilaku transaksional, penggunaan RFM sebagai metode penilaian dan segmentasi saja [9], [17]. Segmentasi pelanggan dengan pendekatan rangking RFM yang efektif juga bisa untuk meningkat kinerja pemasaran [5].

Penelitian lain yang dikembangkan yaitu mencoba untuk mengolah data dengan mengkombinasikan metode analisis RFM dan *phyton* untuk memprediksi perilaku pelanggan dengan merancang strategi pemasaran yang lebih efektif [20]. Adapun cara mendapatkan data yang akurat dan algoritma transaksi dalam waktu dekat, diperlukan uji sampel data yang memiliki permintaan tinggi untuk data yang akurat dan tepat sebagai sarana untuk menguji dan menganalisis RFM melalui metode *machine learning* [3], [15]. Namun, beberapa mengalami kesulitan untuk mendeteksi kapan pelanggan baru berpotensi tersebut akan melakukan transaksi kembali. Ini terutama berlaku untuk hubungan non-kontraktual. Dalam kasus seperti itu, pengelola bisnis harus menetapkan beberapa metrik penting untuk mengidentifikasi pelanggan baru berpotensi yang sebenarnya [14]. Untuk studi kasus penelitian ini mengambil jangka waktu 6 bulan tanpa transaksi untuk menjadi ukuran potensi yang handal. Umumnya, frekuensi, dan nilai transaksi keuangan yang mudah dihitung dan dipahami, tetapi hanya mencakup satu aspek perilaku pelanggan. Untuk mencapai model prediksi berkualitas tinggi, peneliti memerlukan data lengkap tentang kebutuhan pelanggan, opini, karakteristik sosial dan ekonomi, data hubungan, dan lain-lain [11]. Dalam banyak kasus, data semacam itu sulit diambil karena perusahaan kecil dan menengah tidak menerapkan pendekatan sistematis untuk mengumpulkan data tersebut. Sebagian data yang diambil dari data retail untuk

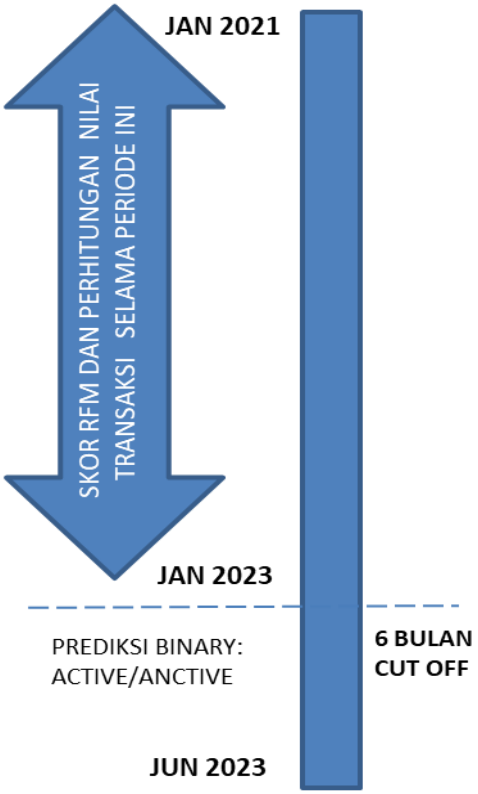
mempermudah prediksi [19]. Salah satu metode yang diterapkan untuk bisa mengklasifikasikan data prediksi yang diharapkan yaitu dengan algoritma *machine learning* sebagai metode yang tepat [4]. Tujuan dari penelitian ini adalah untuk mengusulkan dan mengevaluasi pendekatan prediksi potensi pelanggan baru menggunakan algoritma *machine learning* sebagai analisis data RFM yang dipilih dan dievaluasi dengan variabel masukan yang berbeda.

**Metode**

Data untuk penelitian ini telah diambil dari sistem ERP dari sebuah perusahaan produksi *sheet metal di PT.ABC* yang beroperasi di wilayah Solo, Jawa Tengah. Rentang data adalah dua tahun, dari 2021 hingga 2023 ditunjukkan pada gambar 2. Untuk sampel data tersebut, peneliti mengambil periode *cut off* transaksi yang telah dipilih untuk transaksi sebelum dan sesudah periode tersebut. Periode ini bisa berbeda tergantung pada kekhususan bidang subjek yang akan dihitung. Karena tidak ada kontrak jangka panjang dengan pelanggan, perusahaan tidak dapat mengetahui kapan pelanggan tertentu telah melakukan pergantian sehingga untuk menyelidiki apakah pelanggan telah beralih ke kompetitor atau masih menjadi pelanggan setia, dari data transaksi dapat diklasifikasikan berdasarkan perilaku pelanggan melakukan transaksi kembali. Periode transaksi rata-rata, misal periode dari tanggal pemesanan pertama hingga tanggal pemesanan terakhir untuk semua penjualan adalah 168 hari, dan disarankan untuk memilih 6 bulan sebagai periode pemisahan (lihat gambar 1), tetapi bisa juga 3 bulan, 45 hari, 9 bulan atau 12 bulan, tergantung kebutuhan bisnis.

Untuk setiap pelanggan yang dihitung, sebagai "Jenis Pelanggan" (individu atau perusahaan), "Penghitungan Objek" (jumlah objek untuk setiap pelanggan), "Monetary_Before6mo" (jumlah total transaksi yang dilakukan sebelum 6 bulan sejak tanggal berakhir), "*Recency*_Before6mo" (saat ini sebelum 6 bulan sejak tanggal berakhir), "*Frequency*_6mo" (jumlah transaksi yang dilakukan sebelum 6 bulan sejak tanggal berakhir), "Transactions_after6mo"(jumlah transaksi yang dilakukan dalam 6 bulan terakhir), "Amount_after6mo" (jumlah transaksi yang dilakukan dalam 6 bulan terakhir).

Skor yang dihasilkan, *Recency, Frequency, dan monetary* telah dihitung dalam Software Statistics SPSS dengan fitur analisis RFM-nya. Skor RFM dihitung menggunakan *K-means clustering algoritma* untuk metode klasik ini [13].




**Gambar 1.** Periode cut-off 6 bulan

Penerapan algoritma *machine learning* dilakukan di *Python* dan *PySpark* [12] [16]. Untuk penelitian ini, akan menggunakan kumpulan data transaksi dari ERP yang dipakai untuk mencakup informasi transaksi setiap pelanggan dari seluruh Indonesia. Perhitungan dan segmentasi pelanggan ditunjukkan pada gambar 2 dengan langkah-langkah sebagai berikut :

1. Mempersiapkan data
2. Menentukan periode untuk metrik RFM
3. Menghitung skor RFM
4. Menghitung skor total RFM
5. Segmentasi pelanggan berdasarkan skor
6. Klasifikasi pelanggan berdasarkan segmen dan kreteria
7. Visualisasi hasil perhitungan.

## Hasil dan Pembahasan

*Hasil Penelitian*

Beberapa percobaan telah dilakukan dengan variabel input yang berbeda untuk mengeksplorasi kesesuaian model untuk memprediksi apakah pelanggan akan melakukan transaksi dalam 6 bulan ke depan berdasarkan kumpulan variabel input yang berbeda. Variabel input diambil dari data informasi transaksi setiap pelanggan seperti id_pelanggan, no faktur, tanggal tranksaksi penjualan, deskripsi produk, jumlah penjualan, dan total biaya transaksi penjualan. Berikut tampilan data transaksi yang digunakan untuk pengambilan dataset tersebut dari lima data teratas dan terbawah.

**Tabel 1**. Dataset Faktur Penjualan PT.ABC

PT. ABC SOLO

Sales Invoice by Invoice No

Dari 01 Jan 2021 ke 31 Dec 2023

| Invoice No. | Invoice Da | Quantity | Unit Price | Amount | Item Description | No. Pelanggan |
|---|---|---|---|---|---|---|
| 21/Rek/0128-W# | 18 Jan 2021 | 1,00 | 5.370.000,00 | 5.370.000,00 | Product WF non standard | 1-0159 |
| 21/Rek/0727-W# | 14 Apr 2021 | 11,00 | 1.932.727,30 | 21.260.000,30 | FILING CABINET 4D | 1-0159 |
| 21/Rek/0727-W# | 14 Apr 2021 | 5,00 | 1.932.727,30 | 9.663.636,50 | FILING CABINET 4D | 1-0159 |
| 21/Rek/0727-W# | 14 Apr 2021 | 10,00 | 1.932.727,30 | 19.327.273,00 | FILING CABINET 4D | 1-0159 |
| 21/Rek/0727-W# | 14 Apr 2021 | 2,00 | 1.932.727,30 | 3.865.454,60 | FILING CABINET 4D | 1-0159 |
| ... | ... | ... | ... | ... | ... | ... |
| 23/Rek/0737-M | 17 Mei 2023 | 2,00 | 25.000,00 | 50.000,00 | Harden | 1238 |
| 23/Rek/0739-M | 17 Mei 2023 | 1,00 | 700.000,00 | 700.000,00 | Harden | 1-0399 |
| 23/Rek/0739-M | 17 Mei 2023 | 1,00 | 220.000,00 | 220.000,00 | Harden | 1-0399 |
| 23/Rek/0739-M | 17 Mei 2023 | 1,00 | 120.000,00 | 120.000,00 | Harden | 1-0399 |
| 23/Rek/0736-M | 17 Mei 2023 | 5,00 | 25.000,00 | 125.000,00 | Harden | 1-0112 |

Ada tiga faktor utama Analisis RFM (*Recency, Frequency, dan Monetery*) yang perlu dihitung yaitu :

a. *Recency*: Transaksi terdekat untuk pelanggan baru yang melakukan pembelian.
b. *Frequency*: Jumlah traksaksi setiap pembelian
c. *Monetery:* Total biaya yang dikeluarkan.

Perhitungan ketiga faktor utama ini dengan mengelompokkannya berdasarkan pelanggan dan mengambil "2023/05/17" sebagai tanggal akhir referensi karena ini adalah tanggal transaksi terakhir yang tercantum dalam kumpulan data kami.

Perhitungan *Recency* diperoleh dari kolom *Invoice Date* pada tabel dataset yang diambil dari invoice date terakhir. Dengan menggunakan fungsi F.max() tanggal transaksi terbesar atau terbaru dapat diperoleh. Kemudian nilai *Recency* akan didapatkan dengan mengitung selisih dari tanggal transaksi terakhir dikurangi tanggal transaksi terbaru.

```
# Recency   = Transaksi terdekat untuk pelanggan baru
# Frequency = Jumlah transaksi
# Monetary = Total biaya yang dikeluarkan

# Set 2011/05/17 transaksi terdekat.
# Pelanggan baru yang dihitung per hari.
latest_date = F.to_date(F.lit("2011/05/17"),
'yyyy/MM/dd')

# Perhitungan RFM
# MENGHITUNG RECENCY
day = '2023-05-17'
day = pd.to_datetime(day)
df['date'] - pd.to_datetime (df['date'])
recency = (rtl_data.groupBy("CustomerID").
agg((F.datediff(latest_date,
F.max(F.col("InvoiceDate")))).alias("Recency")))
```

**Gambar 2.** Langkah Perhitungan *Recency*

Hasil nilai *Recency* ditunjukkan dalam gambar 2 dan tabel 2.

**Tabel 2**. Hasil Perhitungan *Recency*

| No. | CUSTOMER ID | LAST INVOICEDATE | *RECENCY* |
|---|---|---|---|
| 1 | 1-3130 | 17-05-2023 - 25-06-2022 | 326 |
| 2 | 1-2544 | 17-05-2023 - 14-05-2023 | 3 |
| 3 | 3-0245 | 17-05-2023 - 16-05-2023 | 1 |
| 4 | 1-0035 | 17-05-2023 - 13-05-2023 | 4 |
| 5 | 1-0002 | 17-05-2023 - 13-05-2023 | 4 |

Langkah selanjutnya menghitung nilai *Frequency* yang didapat dari data *invoice date* untuk mengukur seberapa sering pelanggan melakukan transaksi dalam periode yang ditentukan (misal, 6 bulan terakhir). Disini menggunakan fungsi F.count() bisa memperoleh jumlah transaksi keseluruhan untuk nilai *Frequency*. Hasil dari perhitungan *Frequency* telah ditunjukkan di gambar 3 dan tabel 3.

```
# MENGHITUNG FREQUENCY
frequency = (
rtl_data.groupBy("CustomerID", "InvoiceNo").count()
.groupBy("CustomerID")
.agg(F.count("*").alias("Frequency"))
)
```

**Gambar 3.** Langkah Perhitungan *Frequency*

**Tabel 3**. Hasil Perhitungan *Frequency*

| No. | CUSTOMER ID | *FREQUENCY* |
|---|---|---|
| 1 | 1-3130 | 1 |
| 2 | 1-2544 | 103 |
| 3 | 3-0245 | 4596 |
| 4 | 1-0035 | 199 |
| 5 | 1-0002 | 59 |

Langkah terakhir menghitung nilai *monetary* yang diambilkan dari data kolom *Amount* pada dataset yang berisikan nilai total transaksi penjualan. Dengan menggunakan fungsi F.sum() maka dapat mengitung totalnya yang merupakan nilai angka *monetary* yang dihasilkan. Hasil nilai monetary ini ditunjukkan pada gambar 4 dan tabel 4.

```
# MENGHITUNG MONETARY
monetary = (
rtl_data.groupBy("CustomerID")
.agg(F.sum(F.col("TotalAmount")).alias("Monetary"))
)
```

**Gambar 4.** Langkah Perhitungan Monetary

**Tabel 4**. Hasil Perhitungan Monetary

| No. | CUSTOMER ID | MONETARY |
|---|---|---|
| 1 | 1-3130 | 771.840.000 |
| 2 | 1-2544 | 420.700.000 |
| 3 | 3-0245 | 339.660.000 |
| 4 | 1-0035 | 41.150.000 |
| 5 | 1-0002 | 9.460.000 |

Setelah masing-masing nilai RFM telah didapatkan, langkah selanjutnya menggabungkan ketiga kolom dalam satu kerangka data. Penggabungan ini menggunakan fungsi F.merge untuk menampilkan data *Recency*, *Frequency* dan *Monetary* secara berdampingan seperti yang ditunjukkan pada gambar 5 dan tabel 5.

**Tabel 5**. Hasil klasifikasi pelanggan RFM
Sumber: Dataset PT.ABC

| No. | CUSTOMER ID | R | F | M |
|---|---|---|---|---|
| 1 | 1-3130 | 326 | 1 | 771.840.000 |
| 2 | 1-2544 | 3 | 103 | 420.700.000 |
| 3 | 3-0245 | 1 | 4596 | 339.660.000 |
| 4 | 1-0035 | 4 | 199 | 41.150.000 |
| 5 | 1-0002 | 4 | 59 | 9.460.000 |

```
# MENGGABUNGKAN KETIGA KOLOM RFM MENJADI SATU KERANGKA
df_rfm = df_recency.merge (df_frequency, on = 'CustomerID'),
merge (df_monetary, on = 'CustomerID'). drop (columns = 'invoicedate')
df_rfm.head()
```

**Gambar 5.** Menggabungkan RFM dalam sebuah kerangka data

Eksperimen selanjutnya dapat dilakukan untuk membandingkan apakah algoritma yang dipilih akan memprediksi data pelanggan dengan periode pemisahan yang berbeda, misalnya 6 bulan terakhir. Dari data hasil klasifikasi RFM tersebut langkah selanjutnya menghitung skor RFM.

```
import seaborn as sns

rfm_scores_df = rfm_scores.toPandas()

fig, ax = plt.subplots(1, 3, figsize=(16, 8))

# Recency distribution plot
sns.histplot(rfm_scores_df['Recency'],
kde=True, ax=ax[0])

# Frequency distribution plot
sns.histplot(rfm_scores_df
.query('Frequency < 25')['Frequency'],
kde=True, ax=ax[1])

# Monetary distribution plot
sns.histplot(rfm_scores_df
.query('Monetary < 10000')['Monetary'],
kde=True, ax=ax[2])
```

**Gambar 6.** Experimen Prediksi dengan Pandas Seaborn

Uji coba dilakukan menggunakan pembagian 3 klasifikasi RFM dengan skor segmentasi gabungan, seperti yang ditunjukkan pada gambar 6. Pada awalnya akan diberikan skor spesifik kepada setiap pelanggan untuk masing-masing *Recency*, *Frequency*, dan *Monetery*. Kemudian menggabungkan skor individual tersebut menjadi skor segmentasi gabungan. Setelah itu membagi pelanggan menjadi tiga pembobotan yang sama (33% di setiap bagian) dan memberikan skor dari 1 hingga 3 (terbaik hingga terburuk) untuk setiap bagian. Pembagian kelompok skor 1-3 berdasarkan klasifikasi RFM ditunjukkan di tabel 6.

Untuk *Recency*, kami akan memberikan skor "1" kepada pelanggan yang baru saja membeli (33% pertama), skor "2" untuk kelompok menengah, dan "3" untuk

kelompok ketiga (pelanggan yang terakhir yang membeli sejak lama).

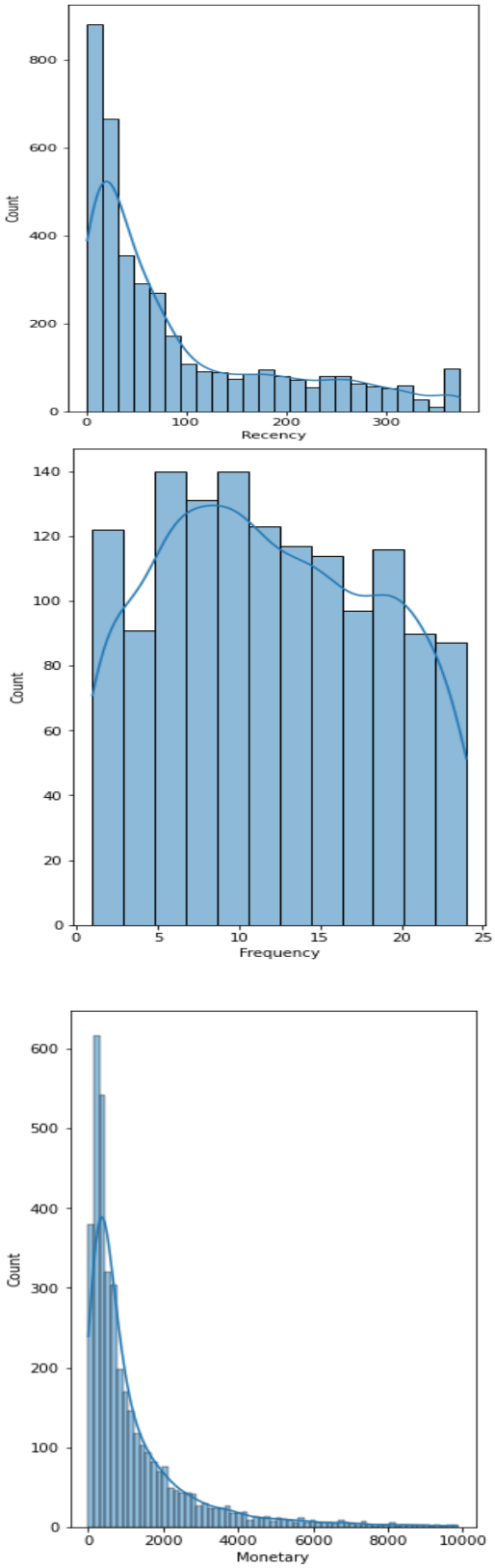
Karena pelanggan yang baru saja membeli memiliki kemungkinan lebih besar untuk melakukan transaksi kembali, skor yang diberikan ke pelanggan lebih baik. Namun untuk *Frequency* dan *Monetery*, kami akan memberikan skor "1" kepada 33% pelanggan terakhir, yaitu pelanggan yang berbelanja lebih sering dan membelanjakan lebih banyak. Dan ditetapkan "3" untuk 33% pelanggan pertama yang lebih jarang berbelanja dan membelanjakan lebih sedikit. Tabel pembagian skor berdasarkan segmen dan kreteria ini ditunjukkan pada tabel 7 dan hasil perhitungan didapatkan dari rumusan pada gambar 11. Pembobotan 33% ini digunakan oleh perusahaan untuk mengidentifikasi dan memberikan skala prioritas pelanggan yang memiliki kontribusi finansial terbesar dengan tetap mempertimbangkan frekuensi dan terkinian transaksi. Pada akhirnya, kita akan mendapatkan nilai segmentasi mulai dari "111" hingga "333" (terbaik hingga terburuk), dan skor gabungan mulai dari 3 hingga 9 seperti yang ditunjukkan pada tabel 8.

**Tabel 6**. Daftar Score RFM dalam pembagian kelompok 33%

| R_SCORE | KELOMPOK 33% |
|---|---|
| 1 | Terbaik transaksi terbaru |
| 2 | Menengah |
| 3 | Terakhir transaksi terlama |

| F_SCORE | KELOMPOK 33% |
|---|---|
| 1 | Transaksi sering dan banyak |
| 2 | Menengah transaksi jarang |
| 3 | Transaksi paling sedikit |

| M_SCORE | KELOMPOK 33% |
|---|---|
| 1 | Nilai transaksi terbesar |
| 2 | Nilai transaksi sedang |
| 3 | Nilai transaksi terkecil |



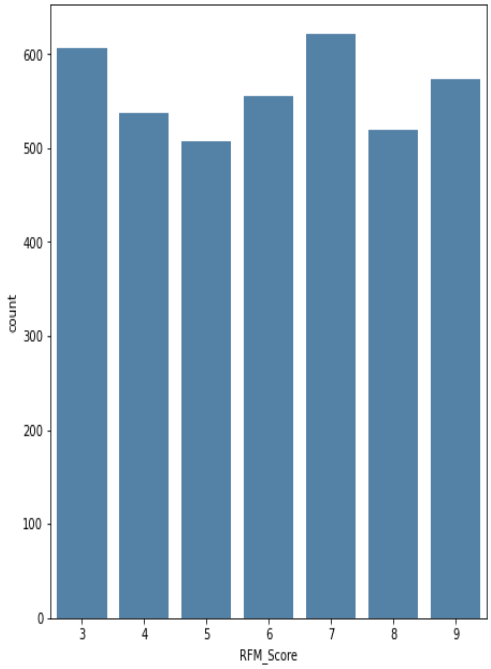

**Gambar 7.** Grafik Segmentasi RFM berdasarkan jumlah pelanggan

**Tabel 7**. Klasifikasi pelanggan berdasarkan segmen dan kreteria

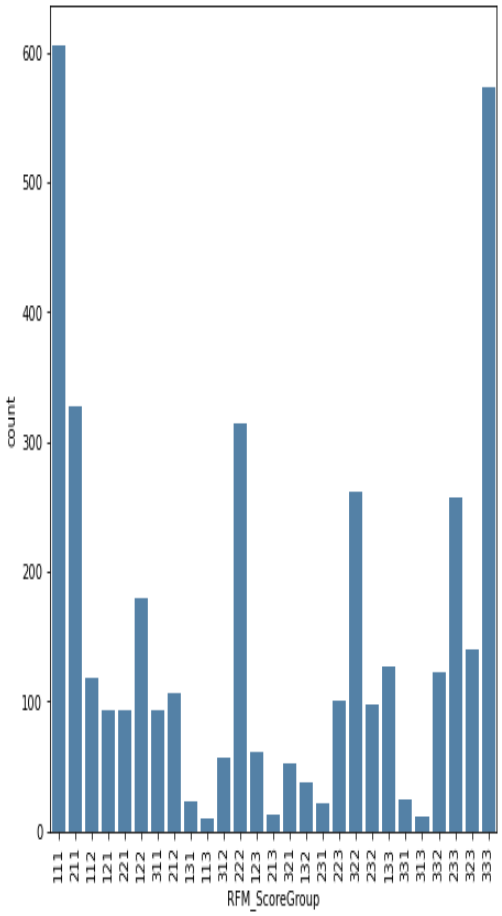
| Segment<br>Kreteria | Awal<br>33% | Tengah<br>33% | Akhir<br>33% |
|---|---|---|---|
| *Recency* | **1**<br>Pembelian 1 bulan | **2**<br>Pembelian 1-3 bulan | **3**<br>Pembelian > 3 bulan |
| *Frequency* | **3**<br>< 25 transaksi | **2**<br>25–70 transaksi | **1**<br>>70 transaksi |
| Monetary | **3**<br>< 5 juta | **2**<br>5-100 juta | **1**<br>>100 juta |

```
# Identify 3 segments RFMScore (ditetapkan 33%)
segments = [0.33, 0.66]
quantiles = rfm_agg_scores.approxQuantile
("RFM_Score", segments, 0)
quantiles

#### Step 1
# Input skor RFM
rfm_scores = (rfm_numbers.withColumn("R_Score",
F.when(F.col("Recency") < quantiles[0][0], F.lit(1))
    .when(F.col("Recency") < quantiles[0][1], F.lit(2))
    .otherwise(F.lit(3)))
    .withColumn("F_Score",
F.when(F.col("Frequency") < quantiles[1][0], F.lit(3))
    .when(F.col("Frequency") < quantiles[1][1], F.lit(2))
    .otherwise(F.lit(1)))
    .withColumn("M_Score",

F.when(F.col("Monetary") < quantiles[2][0], F.lit(3))
    .when(F.col("Monetary") < quantiles[2][1], F.lit(2))
    .otherwise(F.lit(1))))

#### Step 2
# Hitung skor RFM
rfm_agg_scores = (rfm_scores
    .withColumn("RFM_Score", F.col("R_Score") +
F.col("F_Score") + F.col("M_Score"))
    .withColumn("RFM_ScoreGroup",
F.concat(
    F.col("R_Score").cast(StringType()),
F.col("F_Score").cast(StringType()),
    F.col("M_Score").cast(StringType())))
)
```

**Gambar 8.** Hitung Skor berdasarkan kreteria




**Gambar 9.** Skor berdasarkan elemen




**Gambar 10.** Skor terdistribusi

```
# Identify 3 RFMScore (ditetapkan 33%)
segments = [0.33, 0.66]
quantiles = rfm_agg_scores.approxQuantile
("RFM_Score", segments, 0)

# Input 3 level segmentasi
loyalty_level = ['Platinum', 'Gold', 'Silver']

rfm_loyalty = (rfm_agg_scores
    .withColumn("Loyalty",
                F.when((F.col("RFM_Score") <=
                 quantiles[0]), F.lit(loyalty_level[0]))
   when((F.col("RFM_Score") <= quantiles[1]),
                F.lit(loyalty_level[1]))
                .otherwise(F.lit(loyalty_level[2]))))
)
```

**Gambar 11.** Perhitungan segmentasi pelanggan berdasarkan skor

Seperti yang bisa kita lihat di grafik gambar 10 ini data yang terdistribusi skor RFM pelanggan, kita mendapatkan plot yang terdistribusi cukup merata. Namun ketika melihat skor berdasarkan elemen (grafik gambar 9), terlihat bahwa skor tersebut tidak terdistribusi secara merata yang disebabkan hasil proses agregasi. Misalnya, skor rfm "5" (di bagan kiri) dapat dicapai dengan skor elemen "212" atau "131" atau "221" atau bahkan "113". Tapi keduanya tidak sama.

Sejauh ini untuk menghitung *Recency, Frequency, dan Monetery* masing-masing pelanggan r_score, f_score, dan m_score terpisah, dan terakhir rfm-score gabungan.

Keputusan segmentasi yang dipilih bergantung pada kebutuhan bisnis, sehingga dapat dibagi basis pelanggan sesuai keinginan. Namun, untuk mempermudah membagi basis pelanggan menjadi 3 segmen berdasarkan perhitungan skor RFM dan menetapkan indikator loyalitas. Misalkan kita berikan 3 level indikator Platinum, Gold, Silver dengan pengelompokkan :

- Segmen 1 (Platinum): 33% pertama
- Segmen 2 (Gold): 33% — 66%
- Segmen 3 (Silver): 33% terakhir

**Tabel 8**. Hasil RFM pelanggan berdasarkan segmen dan kreteria

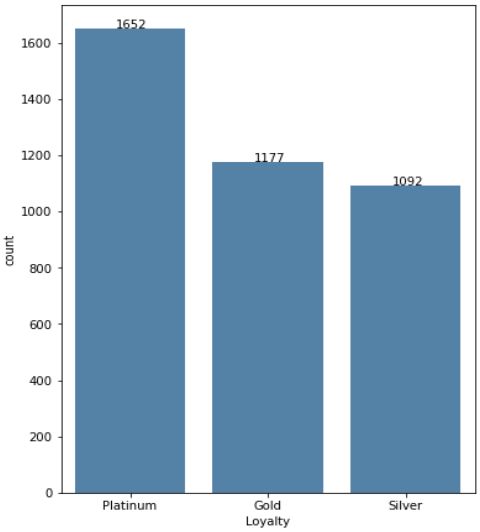
| No | Cust ID | R | F | M (Rp) |
|---|---|---|---|---|
| 1 | 3-1022 | 51 | 30 | 5.988.300 |
| 2 | 1-2954 | 107 | 84 | 14.319.300 |
| 3 | 1-0159 | 17 | 302 | 51.789.600 |
| 4 | 1-2945 | 1 | 213 | 318.336.800 |
| 5 | 1-0612 | 331 | 9 | 1.551.700 |

| No | R_Score | F_Score | M_score |
|---|---|---|---|
| 1 | 2 | 2 | 2 |
| 2 | 3 | 1 | 1 |
| 3 | 1 | 1 | 1 |
| 4 | 1 | 1 | 1 |
| 5 | 3 | 3 | 3 |

| RFM_ Score | RFM_ ScoreGroup | Loyalty |
|---|---|---|
| 6 | 222 | Gold |
| 5 | 311 | Platinum |
| 3 | 111 | Platinum |
| 3 | 111 | Platinum |
| 9 | 333 | Silver |

*Pembahasan*

Dengan memeriksa 3 tingkatan loyalitas diatas maka dapat kita lihat hasil segmentasi menggunakan *Pandas* dan *Seaborn* menghasilkan diagram distribusi sebagai berikut :



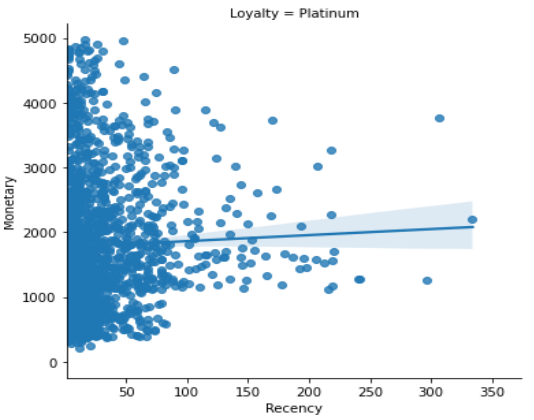

**Gambar 12.** Hasil Segmentasi menggunakan diagram distribusi

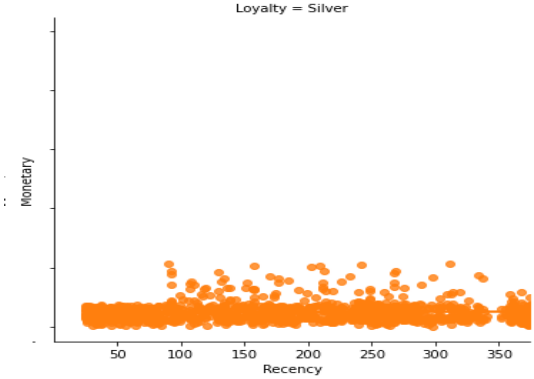
Dari hasil segmentasi pelanggan pada gambar 12, menunjukkan grafik *Recency vs moneter* dan *Recency vs Frequency*, terdapat perbedaan yang jelas antara pelanggan Platinum dan Silver. Pelanggan Platinum cenderung berbelanja lebih sering dan lebih banyak dibandingkan pelanggan lain (kategori pelanggan bagus). Dan pelanggan Silver mungkin sudah kehilangan minat berbelanja sehingga lebih jarang berbelanja dan lebih sedikit (kategori pelanggan tidak terlalu bagus). Namun, pelanggan Gold berada di tengah-tengah masuk dalam kategori pelanggan cukup bagus karena ada lebih banyak aktivitas di antara pelanggan Gold di akhir-akhir transaksi ini.

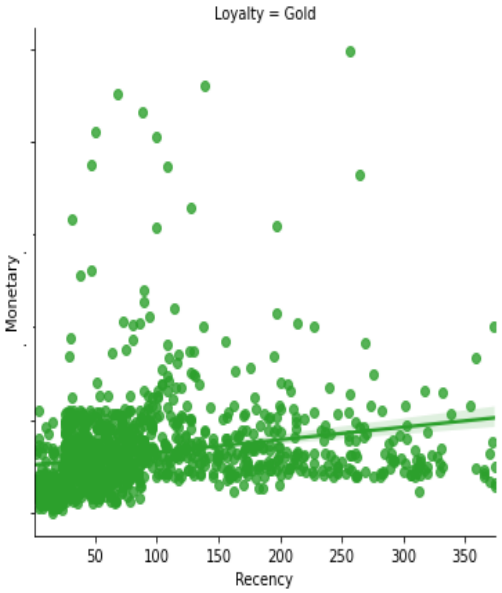
Untuk melihat secara lebih optimal dari bentuk pola sebaran (*scatter plot*) dari ketiga segmen diatas dapat dilihat pada gambar 13-15, batasan *dataset* ditentukan sebagai berikut :

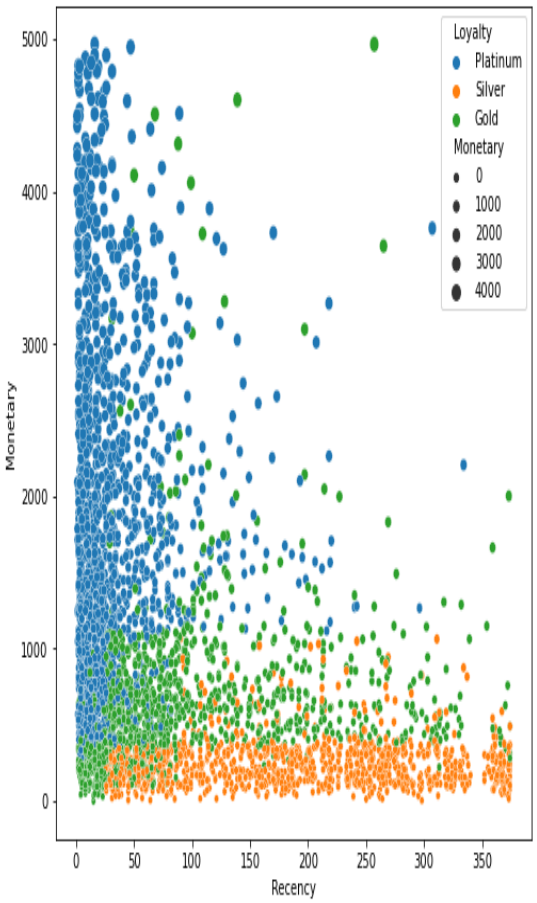
- *Monetary* < 5000 dan *Recency* < 350

<AxesSubplot:xlabel='Monetary', ylabel='*Frequency*'>



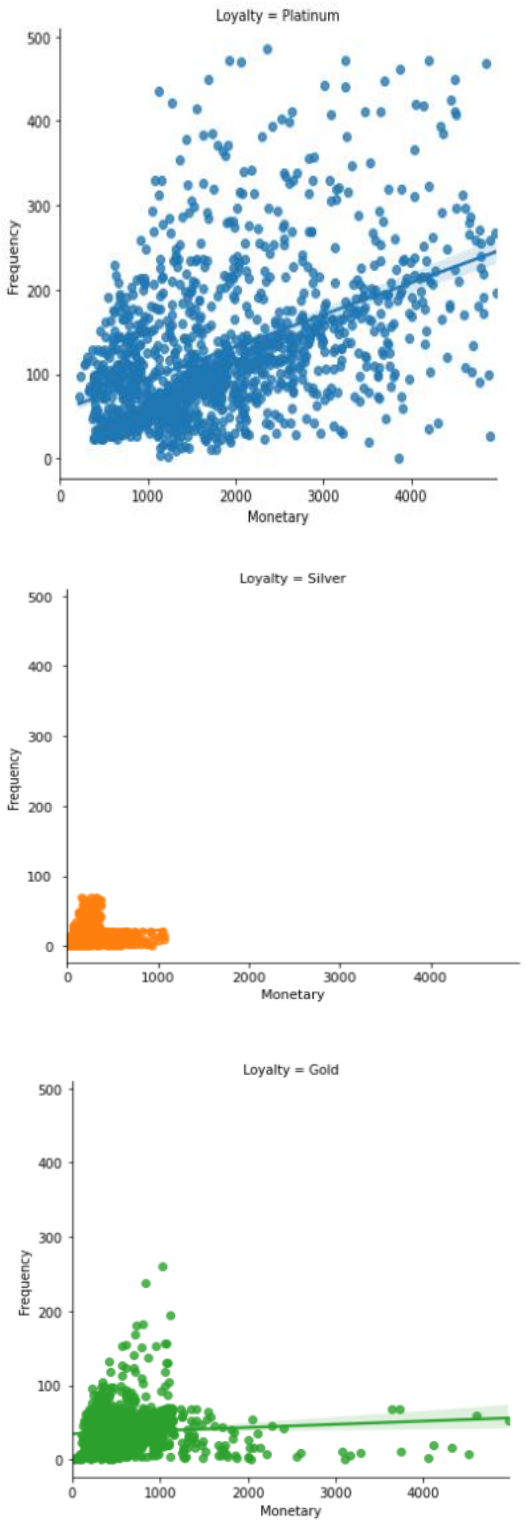




**Gambar 13.** Grafik *Scatter Plot Monetary* vs *Recency*

- *Frequency < 500 dan Recency < 350*

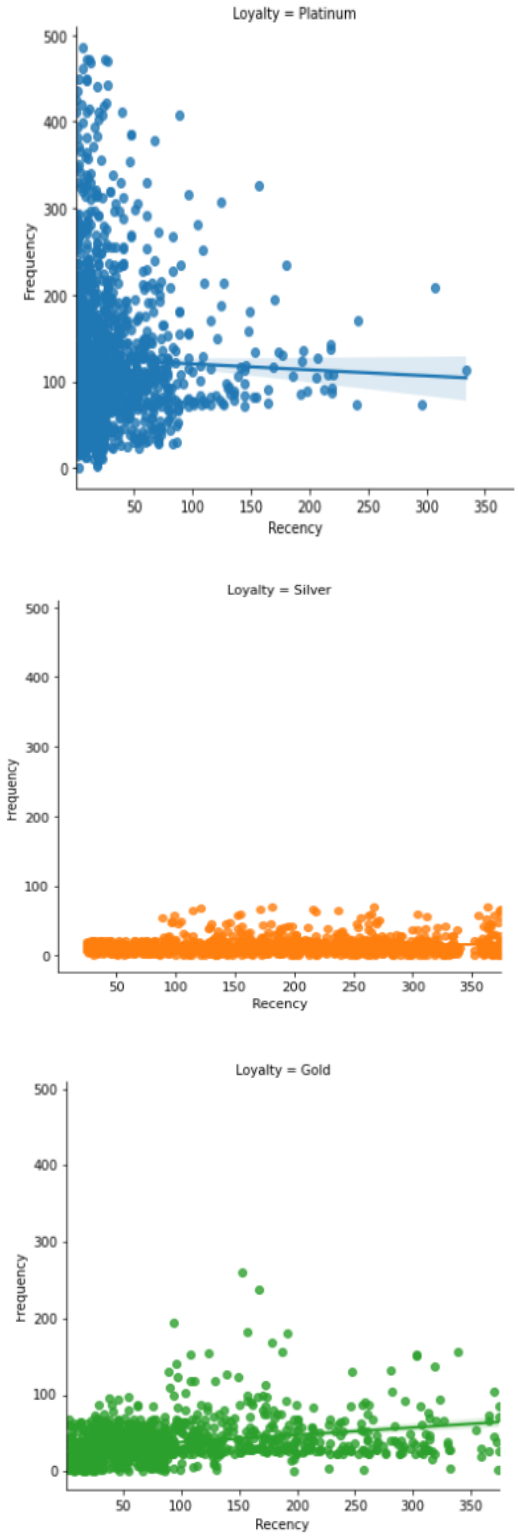
<AxesSubplot:xlabel='Monetary', ylabel='*Frequency*'>




**Gambar 14.** Grafik Scatter Plot *Frequency* vs *Recency*

- *Monetary* < 5000 dan *Frequency* < 500

<AxesSubplot:xlabel='Monetary', ylabel='*Frequency*'>




**Gambar 15.** Grafik *Scatter Plot Monetary* vs *Frequency*

Hasil penelitian ini memberikan implikasi praktis yang signifikan, khususnya bagi PT. ABC dalam meningkatkan efektifitas pengelolaan pelanggan dengan nilai kontribusi tinggi (Platinum) serta merancang strategi retensi untuk pelanggan dengan potensi lebih rendah (Silver) Selain itu melalui pengelolaan pendapatan perusahaan dapat memfokuskan 80% anggaran pemasaran untuk pelanggan Platinum dan Gold yang memberikan kontribusi besar terhadap pendapatan perusahaan. Sedang 20% anggaran pemasaran untuk pelanggan Silver menggunakan metode hemat biaya, seperti mempromosikan dengan berbasis media sosial atau digital marketing. Selain itu, penggunaan *machine learning* dengan *PySpark* menunjukkan bahwa transformasi digital berbasis *big data* dapat meningkatkan efesiensi operasional perusahaan.

## Simpulan

Semua eksperimen mengungkapkan bahwa *Recency, Frequency,* dan *monetary*, baik sebagai nilai absolut maupun sebagai skor, merupakan prediktor kuat dari pencarian pelanggan berpotensi. Dengan algoritma *machine learning*, analisis RFM tidak hanya dapat digunakan sebagai metode segmentasi dan penilaian saja tetapi juga untuk tujuan klasifikasi dan prediksi. Evaluasi beberapa algoritme *machine learning* yang paling banyak digunakan di Phyton menunjukkan bahwa hanya dengan nilai *Recency, Frequency,* dan *moneter* sebagai nilai input, perusahaan dapat memprediksi dengan dasar yang cukup baik apakah pelanggan akan melakukan pembelian dalam waktu yang diinginkan tiap periode. Setelah pemilihan variabel masukan tambahan yang dalam studi kasus terjadi jumlah objek dan jenis pelanggan. Dengan menambahkan variabel masukan lain yang relevan tergantung pada data area subjek, peneliti dapat menyempurnakan proses prediksi yang menghasilkan perkiraan pelanggan berpotensi yang lebih akurat. Di sisi lain, ini akan memberikan keunggulan kompetitif bagi pengembangan penelitian yaitu strategi hubungan pelanggan untuk mempertahankan pelanggan mereka yang menguntungkan dan mencegah persaingan yang tidak diinginkan.

Hasil evaluasi menunjukkan bahwa bagaimana dapat menerapkan model segmentasi manajerial RFM dengan *PySpark* bekerja lebih baik dengan *Recency, Frequency, dan moneter* yang diambil sebagai variabel kontinue. Penelitian di masa mendatang harus dilakukan untuk mengevaluasi kinerja *algoritme machine Learning* yang sama untuk memprediksi pelanggan berpotensi untuk periode pembagian waktu yang berbeda.

## Daftar Pustaka

[1] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means Algorithm: A Comprehensive Survey and Performance Evaluation", Electronics, vol. 9, no. 8, p. 1295, 2020, doi: https://doi.org/10.3390/electronics9081295

[2] A.N. Alrawi and N. Ajlouni, "Intelligent Machine Learning Customer Segmentations Algorithm", MJAIAS, vol. 3, no. 1, May 2022.

[3] D. Bratina and A. Faganel, "Using Supervised Machine Learning Methods for RFM Segmentation: A Casino Direct Marketing Communication Case", Market-Tržište, vol. 35, no.1, pp. 7-22, 2023.

[4] P. Chaudhary, V. Kalra, and S. Sharma, "A hybrid machine learning approach for customer segmentation using rfm analysis", in International Conference on Artificial Intelligence and Sustainable Engineering: Select Proceedings of AISE, pp. 87-100, April 2022.

[5] A. J. Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, "RFM Ranking – an Effective Approach to Customer Segmentation," Journal of King Saud University - Computer and Information Sciences, vol. 33, no. 10, Sep. 2018, doi: https://doi.org/10.1016/j.jksuci.2018.09.004

[6] A. Dewabharata, "Customer Segmentation Using the K-Means Clustering as a Strategy to Avoid Overstock in Online Shop Inventory", Proceedings of the 1st International Conference on Contemporary Risk Studies (ICONIC-RS), South Jakarta,

DKI Jakarta, March-1 April 2022, doi: https://10.4108/eai.31-3-2022.2320688

[7] R. Gustriansyah, N. Suhandi, and F. Antony, “Clustering optimization in RFM analysis based on k-means”, Indonesian Journal of Electrical Engineering and Computer Science, vol. 18, no.1, pp. 470-477, 2022.

[8] J. Joung and H. Kim, “Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews”, International Journal of Information Management, vol. 70, p. 102641, Jun. 2023, doi: https://doi.org/10.1016/j.ijinfomgt.2023.102641

[9] J. Ma, “E-commerce Customer Segmentation Based on RFM Model”, Lecture Notes in Electrical Engineering, pp. 926–931, 2022, doi: https://doi.org/10.1007/978-981-16-8052-6_118

[10] I. Maryani, D. Riana, R. D. Astuti, A. Ishaq, Sutrisno, and E. A. Pratama, “Customer Segmentation based on RFM model and Clustering Techniques With K-Means Algorithm”, 2018 Third International Conference on Informatics and Computing (ICIC), Oct. 2018, doi: https://doi.org/10.1109/iac.2018.8780570

[11] S. Monalisa, Y. Juniarti, E. Saputra, F. Muttakin, and T.K. Ahsyar, “Customer segmentation with RFM models and demographic variable using DBSCAN algorithm”, Telkomnika (Telecommunication Computing Electronics and Control), vol. 21, no. 4, pp. 742-749, 2023, doi: https://doi.org/10.12928/TELKOMNIKA.v21i4.22759

[12] S. Ozan, “A Case Study on Customer Segmentation by using Machine Learning Methods”, 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Sep. 2018, doi: https://doi.org/10.1109/idap.2018.8620892

[13] P. Saurabh, H. Khan, S. Mehta and U. Mandawkar, "Study of Customer Segmentation Using K-Means Clustering and RFM Modelling", Journal of Engineering Sciences, vol. 12, no. 6, pp. 556-559, 2021.

[14] M. A. Rahim, M. Mushafiq, S. Khan, and Z. A. Arain, “RFM-based repurchase behavior for customer classification and segmentation”, Journal of Retailing and Consumer Services, vol. 61, p. 102566, Jul. 2021, doi: https://doi.org/10.1016/j.jretconser.2021.102566

[15] İ. SABUNCU, E. TÜRKAN, and H. POLAT, “Customer Segmentation and Profiling with Rfm Analysis,” Turkish Journal of Marketing, vol. 5, no. 1, pp. 22–36, Apr. 2020, doi: https://doi.org/10.30685/tujom.v5i1.84

[16] AH. Sial, S.Y.S. Rashdi, and A.H. Khan, "Comparative analysis of data visualization libraries Matplotlib and Seaborn in Python," International Journal, vol. 10, no. 1, pp. 277-281, 2021.

[17] S. Wan, J. Chen, Z. Qi, W. Gan, and L. Tang, “Fast RFM model for customer segmentation”, in WWW 2022 - Companion Proceedings of the Web Conference, April 2022, doi: https://doi.org/10.1145/3487553.3524707

[18] Z. Xian, P. Keikhosrokiani, C. XinYing, and Z. Li, “An RFM Model Using K-Means Clustering to Improve Customer Segmentation and Product Recommendation”, Handbook of Research on Consumer Behavior Change and Data Analytics in the Socio-Digital Era, pp. 124-145, 2022, IGI-Global, Malaysia, doi: https://10.4018/978-1-6684-4168-8.ch006

[19] F. Yoseph and M. Heikkila, “Segmenting Retail Customers with an Enhanced RFM and a Hybrid Regression/Clustering Method,” IEEE Xplore, Dec. 01, 2018, doi: https://ieeexplore.ieee.org/abstract/document/8614012/

[20] A. M. A. Zamil and T. G. Vasista, “Customer segmentation using RFM analysis: Realizing through Python implementation”, Pacific Business

Review International, vol.13, pp. 24-36,
May 2021.