

PENINGKATAN PROFIT MELALUI STRATEGI DISKON DINAMIS MENGGUNAKAN REINFORCEMENT LEARNING PADA DATA TRANSAKSI RITEL DARING

Andre Hasudungan Lubis^{1*}, Rizki Muliono², Nanda Novita³, Susilawati Susilawati⁴,
Muhathir Muhathir⁵, Sayuti Rahman⁶

^{1,2,3,4,5,6}Program Studi Teknik Informatika, Fakultas Teknik, Universitas Medan Area, Indonesia

e-mail: ¹andrehasudunganlubis@uma.ac.id, ²rizkimuliono@staff.uma.ac.id,
³nandanovita@staff.uma.ac.id, ⁴susilawati@staff.uma.ac.id,
⁵muhathir@staff.uma.ac.id, ⁶sayutirahman@staff.uma.ac.id

ABSTRAK

Strategi diskon memegang peranan penting dalam profitabilitas e-commerce, namun sebagian besar platform masih mengandalkan tarif diskon tetap yang tidak memperhitungkan perilaku pelanggan yang beragam. Penelitian ini mengusulkan kerangka optimasi diskon dinamis berbasis algoritma Q-Learning, di mana keputusan pemberian diskon dimodelkan sebagai *Markov Decision Process* (MDP) dengan ruang keadaan (*state space*) yang dibangun dari dua indikator perilaku pelanggan, yaitu pendapatan transaksi dan frekuensi pembelian. Setiap transaksi pelanggan dipetakan ke salah satu dari enam keadaan diskrit yang terbentuk dari kombinasi tiga kategori pendapatan (*Low, Medium, High*) dan dua kategori frekuensi (*Low, High*), dan agen memilih salah satu dari lima tingkat diskon (0%, 5%, 10%, 15%, 20%) untuk memaksimalkan profit kumulatif. Mekanisme elastisitas harga disematkan dalam fungsi reward untuk mensimulasikan respons permintaan terhadap perubahan harga. Kerangka ini dievaluasi pada dataset Online Retail II yang terdiri dari 449.107 transaksi yang telah dibersihkan. Hasil penelitian menunjukkan bahwa Q-Learning menghasilkan total profit sebesar 10.432.467,30, mengungguli seluruh lima baseline diskon tetap, dengan peningkatan profit masing-masing sebesar +10,77%, +16,60%, +23,08%, +30,32%, dan +38,46%. Kebijakan yang dipelajari menunjukkan pola diskon yang adaptif pada keenam state, dengan pilihan optimal berupa diskon 0%, 10%, 15%, dan 20%, sementara diskon 5% tidak menjadi pilihan optimal pada state mana pun. Hal ini menunjukkan bahwa besaran diskon tidak hanya ditentukan oleh frekuensi pembelian, tetapi juga mempertimbangkan tingkat pendapatan pelanggan serta trade-off antara peningkatan permintaan dan margin keuntungan. Temuan ini menunjukkan bahwa pendekatan berbasis RL menawarkan alternatif yang lebih tepat dan mampu meningkatkan profit secara kompetitif dibandingkan dengan strategi promosi yang bersifat statis.

Kata kunci: Optimasi Diskon Dinamis; Reinforcement Learning; Q-Learning; E-commerce; Elastisitas Harga.

ABSTRACT

Discount strategies play a critical role in e-commerce profitability, yet most platforms continue to rely on fixed discount rates that fail to account for heterogeneous customer behavior. This study proposes a dynamic discount optimization framework based on the Q-Learning algorithm, in which discount decisions are modeled as a Markov Decision Process (MDP) with a state space constructed from two customer behavioral indicators: transaction revenue and purchase frequency. Each customer transaction is mapped to one of six discrete states formed by combining

three revenue categories (Low, Medium, High) and two frequency categories (Low, High), and the agent selects from five discount levels (0%, 5%, 10%, 15%, 20%) to maximize cumulative profit. A price elasticity mechanism is incorporated into the reward function to simulate demand responsiveness to price reductions, enabling the agent to learn non-trivial discount preferences. The framework is evaluated on the Online Retail II dataset comprising 397,885 cleaned transactions. Experimental results show that the proposed method achieves a total profit of 10,586,828.77, outperforming four of five fixed discount baselines with profit improvements of +18.80%, +11.16%, +5.60%, and +1.65% over the 0%, 5%, 10%, and 15% uniform discount scenarios, respectively. The learned policy consistently assigns 15% discounts to Low-frequency customers and 20% discounts to High-frequency customers, demonstrating that purchase frequency is the primary driver of optimal discount differentiation. These findings confirm that RL offers a more targeted and profit-competitive alternative to static promotional strategies.

Keywords: *Dynamic Discount Optimization; Reinforcement Learning; Q-Learning; E-commerce; Price Elasticity.*

1. PENDAHULUAN

Pertumbuhan industri e-commerce telah meningkatkan persaingan antar pelaku usaha dalam menarik dan mempertahankan pelanggan [1]. Sebagai salah satu strategi utama, pihak bisnis menggunakan kebijakan diskon untuk mendorong volume pembelian, mengingat harga secara langsung mempengaruhi perilaku belanja konsumen [2]. Oleh karena itu, kebijakan diskon menjadi komponen penting dalam strategi peningkatan pendapatan dan loyalitas pelanggan di lingkungan bisnis digital [3]. Meskipun demikian, sebagian besar platform e-commerce masih menetapkan tarif diskon secara tetap tanpa mempertimbangkan variasi tingkat diskon [4]. Pendekatan ini mudah diimplementasikan, namun tidak mampu menangkap fluktuasi pola pembelian pelanggan maupun dinamika pasar yang terus berubah, sehingga kebijakan diskon statis berpotensi tidak optimal dalam semua situasi [5]. Oleh karena itu, diperlukan pendekatan yang lebih adaptif dan dinamis dalam penentuan strategi diskon.

Dalam beberapa tahun terakhir, pengambilan keputusan berbasis data (*data-driven*) telah mendorong inovasi sistem penetapan harga dan promosi yang adaptif terhadap perilaku pelanggan [6]. Reinforcement Learning (RL) muncul sebagai paradigma machine learning yang menjanjikan untuk masalah pengambilan keputusan sekuensial [7]. Berbeda dari metode klasik yang bergantung pada aturan tetap, RL memberdayakan agen untuk belajar melalui interaksi aktif dengan lingkungan dan menerima *reward* sebagai umpan balik atas setiap aksi [8], sehingga cocok untuk optimasi penetapan harga dan diskon secara dinamis. Di antara berbagai teknik RL, Q-Learning merupakan salah satu metode yang paling populer karena memiliki dasar teoretis yang sederhana, biaya komputasi yang relatif rendah, serta tidak bergantung pada pengetahuan tentang dinamika lingkungan saat menentukan kebijakan [9]. Sebagai bagian dari *Model-Free Reinforcement Learning*, Q-Learning menawarkan konvergensi yang lebih baik dan implementasi yang lebih mudah dengan memanfaatkan pengalaman langsung dari agen [9]. Melalui eksplorasi dan eksploitasi iteratif, Q-Learning dapat mengidentifikasi kebijakan diskon optimal untuk memaksimalkan *reward* kumulatif pada berbagai kondisi pelanggan [10], sekaligus memungkinkan penetapan harga personal berdasarkan riwayat transaksi.

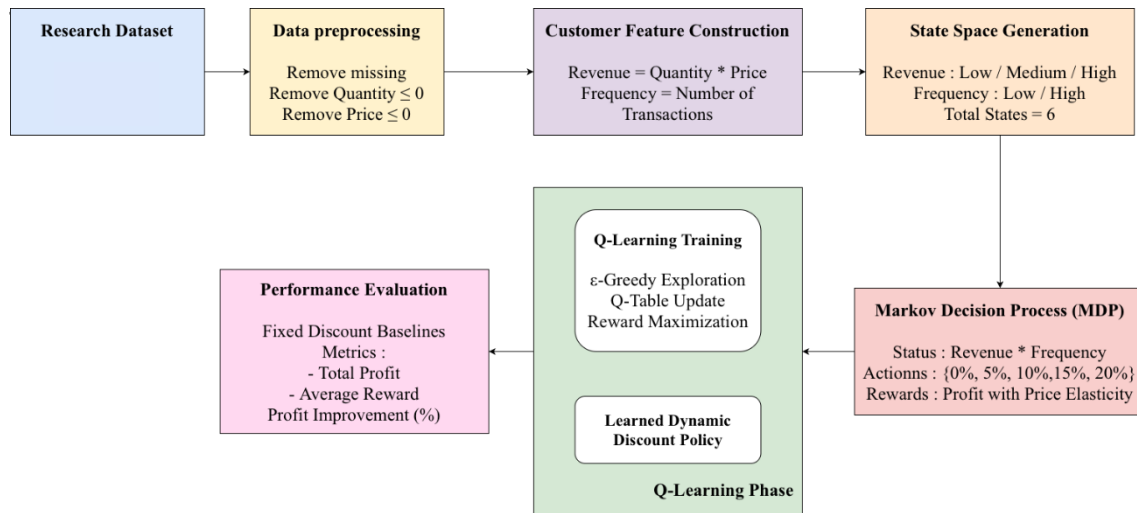
Beberapa penelitian terdahulu telah menerapkan RL dalam konteks penetapan harga. Apte et al. [11] menggunakan Q-Learning untuk penetapan harga dinamis pada sektor ritel, menghasilkan kinerja pendapatan yang lebih unggul dibandingkan metode tradisional. Mullapudi [12] mengembangkan strategi penetapan harga dinamis berbasis Q-Learning yang menyesuaikan diri dengan fluktuasi pasar dan perilaku kompetitif untuk meningkatkan profitabilitas. MuhamedAle et al. [13] mengembangkan algoritma RL yang menyesuaikan harga secara real-time berdasarkan permintaan pasar, perilaku pelanggan, dan harga kompetitor, menggunakan Q-Learning untuk menemukan titik harga optimal melalui trial and error.

Berdasarkan tinjauan tersebut, terdapat dua gap penelitian, yakni belum ada model diskon dinamis berbasis Q-Learning yang mempertimbangkan kombinasi nilai transaksi dan frekuensi pembelian sebagai representasi keadaan (*state*). Selain itu, belum ada evaluasi komparatif terhadap diskon tetap sebagai *baseline*. Oleh karena itu, penelitian ini mengusulkan model optimasi diskon dinamis berbasis Q-Learning pada data transaksi e-commerce, dengan ruang *state* yang menggabungkan kategori nilai transaksi dan frekuensi pembelian, serta mengintegrasikan elastisitas harga ke dalam fungsi *reward* agar model lebih realistis secara ekonomi.

Tujuan penelitian ini dirumuskan dalam tiga aspek utama: (1) mengembangkan model diskon dinamis berbasis Q-Learning yang merepresentasikan kondisi pelanggan melalui kombinasi nilai transaksi dan frekuensi pembelian; (2) mengevaluasi kinerjanya menggunakan dataset Online Retail II; dan (3) membandingkan profit strategi Q-Learning terhadap lima skenario diskon tetap (0%, 5%, 10%, 15%, dan 20%) sebagai *baseline*. Kontribusi utama penelitian ini adalah: (1) model diskon dinamis berbasis Q-Learning yang adaptif terhadap karakteristik transaksi pelanggan; (2) evaluasi komprehensif terhadap lima skenario diskon tetap; dan (3) integrasi elastisitas harga ke dalam fungsi *reward* untuk mensimulasikan respons permintaan secara lebih realistis.

2. METODE PENELITIAN

Penelitian ini mengembangkan strategi berbasis RL untuk optimasi diskon dinamis menggunakan data transaksi dari dataset Online Retail II, melalui enam tahapan utama: (1) pengumpulan dan pembersihan data transaksi; (2) perhitungan indikator perilaku pelanggan (pendapatan transaksi dan frekuensi pembelian) dan konversinya menjadi variabel kategorikal pembentuk *state*; (3) formulasi permasalahan sebagai *Markov Decision Process* (MDP), dengan status pelanggan sebagai *state* dan pilihan diskon sebagai *action*; (4) pelatihan agen Q-Learning untuk memperoleh kebijakan diskon dinamis melalui interaksi dengan ruang keadaan, dengan fungsi *reward* berbasis elastisitas harga; (5) penerapan kebijakan yang telah dipelajari untuk menentukan strategi diskon optimal pada tiap kondisi pelanggan; dan (6) perbandingan kinerja strategi tersebut dengan strategi diskon tetap tradisional melalui evaluasi profitabilitas. Gambar 1 menyajikan arsitektur umum dari penelitian ini.



Gambar 1. Arsitektur Umum Penelitian

2.1. Dataset Penelitian

Penelitian ini menggunakan dataset Online Retail II, sebuah dataset transaksi yang tersedia untuk umum dan diperoleh dari Kaggle dengan tautan (https://www.kaggle.com/datasets/matteo2002/retail-dataset?select=invoice_items.csv). Dataset ini mencatat transaksi komersial dari sebuah toko ritel atau marketplace, yang mencakup periode dari Januari 2014 hingga Desember 2015. Beberapa atribut yang tersedia dalam dataset asli meliputi informasi invoice, identifikasi produk, jumlah transaksi, harga produk, identifikasi pelanggan, dan lokasi pelanggan. Total keseluruhan data adalah sebanyak 449.149 baris. Dalam penelitian ini, hanya atribut yang relevan dengan proses optimasi diskon yang digunakan. Tabel 1 merangkum variabel-variabel yang digunakan selama analisis.

Tabel 1. Atribut yang Digunakan dalam Penelitian

Atribut	Deskripsi
Invoice ID	Identifikasi unik transaksi atau invoice yang digunakan untuk membedakan setiap transaksi penjualan.
Product ID	Identifikasi unik produk yang tercatat dalam transaksi.
Quantity	Jumlah unit produk yang dibeli dalam suatu transaksi.
Price	Harga jual per unit produk (£).
Line Total	Total nilai transaksi untuk suatu item, yang merepresentasikan hasil perkalian antara jumlah produk dan harga.
Item	Nama atau deskripsi produk yang tercatat dalam transaksi.
Category	Kategori produk yang digunakan untuk mengelompokkan item berdasarkan jenis atau klasifikasinya.
Date	Tanggal dan waktu terjadinya transaksi.
Customer ID	Identifikasi unik pelanggan yang melakukan transaksi.
Customer Type	Jenis atau tipe pelanggan berdasarkan klasifikasi yang tersedia pada dataset.

2.2. Tahap Prapemrosesan

Dataset dalam penelitian ini memiliki beberapa data yang tidak lengkap, transaksi yang dibatalkan, serta nilai-nilai yang tidak mencerminkan aktivitas penjualan yang valid, sehingga prapemrosesan data dibutuhkan untuk meningkatkan akurasi dan reliabilitas kinerja model [14]. Sejumlah tahapan prapemrosesan dilakukan untuk meningkatkan kualitas data dan memastikan konsistensi sebelum proses RL dilakukan.

Pertama, transaksi dengan $Quantity \leq 0$ dikeluarkan karena biasanya mewakili retur, penyesuaian, atau kesalahan input data. Kemudian, data dengan $Price \leq 0$ dihapus untuk mengeliminasi barang gratis dan anomali harga yang dapat mengganggu perhitungan pendapatan. Tabel 2 merangkum hasil dari setiap tahap prapemrosesan terhadap dataset.

Tabel 2. Atribut yang Digunakan dalam Penelitian

Langkah	Deskripsi	Jumlah Data
Data awal	Dataset asli	449.149
Hapus $Quantity \leq 0$	Menghapus transaksi retur, penyesuaian, atau kesalahan input	449.149
Hapus $Price \leq 0$	Menghapus barang gratis dan kesalahan harga	449.149
Dataset akhir	Transaksi bersih	449.107

Setelah penghapusan data yang tidak valid, dua atribut level transaksi diturunkan sebagai dasar pembentukan ruang state, yaitu *Revenue* dan *Frequency*. *Revenue* (Rv) dihitung sebagai hasil kali *Quantity* (Qy) dan *Price* (Pr) untuk setiap transaksi:

$$Rv_i = Qy_i * Pr_i \quad (1)$$

Frekuensi pembelian (Fr) dihitung untuk setiap pelanggan sebagai jumlah total *invoice* unik yang terkait dengan Customer ID mereka selama periode observasi. Frekuensi pembelian dihitung sebagai jumlah total transaksi yang terkait dengan seorang pelanggan:

$$Fr_c = \sum_{j=1}^{n_c} 1 \quad (2)$$

Dari Persamaan (2), Fr_c merepresentasikan frekuensi pembelian pelanggan c , dan n_c menunjukkan jumlah transaksi yang dilakukan oleh pelanggan tersebut selama periode observasi. Kedua fitur ini kemudian didiskritisasi menjadi kategori pembentuk state. *Revenue* dibagi menjadi tiga tingkat berdasarkan kuartil:

$$Rv_state \begin{cases} \text{Low,} & Rv < Q_1 \\ \text{Med,} & Q_1 \leq Rv < Q_3 \\ \text{High,} & Rv \geq Q_3 \end{cases} \quad (3)$$

Dimana, nilai Q_1 dan Q_3 adalah kuartil pertama dan ketiga distribusi pendapatan. Frekuensi pembelian dibagi menjadi dua tingkat berdasarkan median:

$$Fr_state \begin{cases} \text{Low,} & Fr < \text{Median}(Fr) \\ \text{High,} & Fr \geq \text{Median}(Fr) \end{cases} \quad (4)$$

Kombinasi dari tiga kategori pendapatan dan dua kategori frekuensi membentuk enam state pelanggan yang selanjutnya digunakan dalam formulasi MDP yang dijelaskan pada bagian berikutnya.

2.3. MDP

Permasalahan optimasi diskon dinamis diformulasikan sebagai MDP, kerangka matematis untuk pengambilan keputusan sekuensial di bawah ketidakpastian [15]. Proses pemberian diskon dimodelkan sebagai interaksi antara agen dan lingkungan, di mana agen belajar memilih tingkat diskon yang memaksimalkan profit kumulatif. MDP didefinisikan oleh tuple (S, A, R, γ) , dengan S mewakili *state space*, A menunjukkan ruang aksi (*action space*), R adalah fungsi *reward*, dan γ adalah faktor diskon yang mengatur bobot *reward* di masa depan relatif terhadap *reward* langsung [16].

State space dibangun dari dua indikator: pendapatan transaksi (tiga tingkat: *Low*, *Medium*, *High*) dan frekuensi pembelian (dua tingkat: *Low*, *High*). Perkalian Cartesian keduanya membentuk enam *state*:

$$S = \{(L, L), (L, H), (M, L), (M, H), (H, L), (H, H)\} \quad (5)$$

Elemen pertama menunjukkan kategori Rv , elemen kedua kategori Fr . Representasi ini memungkinkan agen membedakan, misalnya, pembeli bernilai tinggi yang sering berbelanja versus pembeli bernilai rendah yang jarang berbelanja, sehingga dapat diberikan diskon yang sesuai. Di sisi lain, *action space* A mendefinisikan tarif diskon yang tersedia, mencerminkan strategi promosi umum di ritel [16], dan diformulasikan:

$$A = \{0\%, 5\%, 10\%, 15\%, 20\%\} \quad (6)$$

Setiap *action* $a \in A$ merepresentasikan persentase pengurangan harga per unit; 0% berarti tanpa diskon, nilai lebih tinggi berarti promosi lebih agresif. Fungsi *reward* R dirancang untuk mengukur keuntungan dari keputusan diskon pada setiap transaksi. Saat diskon diberikan, harga per unit turun, namun dapat merangsang permintaan sehingga total pembelian meningkat. Untuk menangkap *trade-off* ini, mekanisme elastisitas harga disematkan ke fungsi *reward*. Jumlah pembelian yang disesuaikan setelah diskon:

$$Qy_i^* = Qy_i * (1 + e * d) \quad (7)$$

Di mana Qy_i^* adalah jumlah yang disesuaikan dengan permintaan (*demand-adjusted quantity*), Qy_i adalah jumlah asli, e adalah koefisien elastisitas harga, dan d adalah tarif diskon yang diterapkan. *Reward* untuk transaksi ke- i di bawah *action* d kemudian dihitung sebagai:

$$R_i(d) = Qy_i * (1 + e * d) * Pr_i * (1 - d) \quad (8)$$

Di mana Pr_i adalah harga per unit dari transaksi tersebut. Koefisien elastisitas e mencerminkan asumsi bahwa diskon 10% meningkatkan permintaan 25%. Formulasi ini menangkap *trade-off non-linear* antara penurunan harga dan stimulasi permintaan, mendorong agen mengidentifikasi tingkat diskon yang memaksimalkan pendapatan, bukan sekadar meminimalkan atau memaksimalkan tarif diskon.

2.4. Algoritma Q-Learning

Q-Learning adalah algoritma RL berbasis *model-free* yang mempelajari fungsi nilai-aksi optimal $Q^*(s, a)$ tanpa memerlukan pengetahuan eksplisit tentang dinamika transisi lingkungan [17]. Algoritma ini menyimpan Q-table, yakni setiap entri $Q(s, a)$ merepresentasikan *reward* kumulatif yang diharapkan dari pengambilan *action* a pada *state* s dan mengikuti kebijakan optimal setelahnya [18]. Melalui interaksi berulang, Q-table diperbarui secara iteratif hingga konvergen. Pada setiap langkah pelatihan, agen mengamati *state* saat ini s , memilih aksi a , menerima *reward* r , dan bertransisi ke *state* berikutnya s' . Nilai Q diperbarui sesuai dengan persamaan Bellman [19]:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[r + \gamma \max_{a'} Q(s', a') - Q(s, a) \right] \quad (9)$$

Dimana nilai $\alpha \in (0,1)$ adalah *learning rate*, nilai $\gamma \in (0,1)$ adalah *discount factor*, nilai r adalah *reward* langsung, dan nilai $\max_{a'} Q(s', a')$ adalah nilai Q maksimum dari *state* berikutnya. Suku dalam kurung disebut temporal difference (TD) error, yang mengukur selisih antara estimasi nilai Q saat ini dengan nilai target [20]. Untuk menyeimbangkan eksplorasi pasangan *state-action* yang belum diketahui dan eksploitasi aksi bernilai tinggi yang sudah diketahui, diterapkan kebijakan ϵ -greedy dengan nilai ϵ (*exploration rate*) tetap sepanjang pelatihan, sehingga tingkat eksplorasi konsisten di seluruh episode [22]. Hal ini mencegah agen konvergen prematur ke kebijakan suboptimal akibat eksplorasi yang tidak memadai pada pasangan *state-aksi* yang jarang dikunjungi.

Pelatihan Q-Learning dilakukan selama sejumlah episode tetap, di mana setiap episode adalah satu putaran lengkap melalui seluruh dataset transaksi. Seluruh nilai Q diinisialisasi nol di awal pelatihan. Pada setiap episode, urutan transaksi diacak untuk mengurangi bias urutan, dan agen memproses setiap transaksi secara berurutan: mengidentifikasi *state*, memilih aksi via kebijakan ϵ -greedy, menghitung *reward*, mengambil sampel *state* berikutnya, dan memperbarui Q-table. Proses ini diulang hingga jumlah episode tercapai. Hyperparameter pelatihan Q-Learning dirangkum pada Tabel 3.

Tabel 3. Konfigurasi Hyperparameter pada Q-Learning

Hyperparameter	Simbol	Nilai
<i>Learning rate</i>	α	0.1
<i>Discount factor</i>	γ	0.9
<i>Exploration rate</i>	ϵ	0.1
Jumlah episode	E	1,000
Koefisien elastisitas harga	e	2.5

2.5. Skenario Eksperimen

Penelitian ini akan melakukan perbandingan dengan lima kebijakan diskon tetap sebagai baseline. Setiap *baseline* tersebut merupakan pendekatan diskon statis tradisional, di mana tarif diskon yang sama diterapkan secara konsisten pada seluruh transaksi tanpa mempertimbangkan karakteristik pelanggan maupun kondisi transaksi yang sedang berlangsung. Skenario eksperimen dapat dilihat pada Tabel 4.

Tabel 4. Skenario eksperimen

Skenario	Metode	Keterangan
1	Diskon Tetap 0%	Tanpa pengurangan harga; menjadi batas atas pendapatan (revenue upper bound) tanpa aktivitas promosi.
2	Diskon Tetap 5%	Strategi promosi konservatif dengan pengurangan harga minimal.
3	Diskon Tetap 10%	Tarif diskon tetap yang paling umum diterapkan dalam praktik ritel; menjadi <i>baseline</i> tradisional utama.
4	Diskon Tetap 15%	Strategi promosi moderat-hingga-agresif, menukar margin lebih besar dengan respons permintaan yang lebih tinggi.
5	Diskon Tetap 20%	Strategi diskon tetap paling agresif; pengurangan harga tertinggi dengan pengorbanan pendapatan per unit terbesar.
6	Q-Learning (Diusulkan)	Tarif diskon ditentukan secara adaptif oleh agen berdasarkan state transaksi, menggunakan kebijakan optimal dari Q-table yang telah konvergen.

Perbedaan utama antara skenario diskon tetap dan metode yang diusulkan terletak pada tingkat adaptabilitasnya. Sementara diskon tetap menerapkan tarif sama untuk semua transaksi, agen Q-Learning menyesuaikan diskon berdasarkan pendapatan dan pola pembelian pelanggan, sehingga strategi lebih terarah dan menguntungkan.

2.5. Metrik Evaluasi

Dalam penelitian ini, kinerja strategi diskon berbasis Q-Learning dan skenario diskon dievaluasi menggunakan dua metrik profitabilitas. Pertama, *Total Profit* (TP) digunakan untuk menunjukkan keuntungan kumulatif dari semua transaksi dan menjadi indikator utama keberhasilan strategi dalam memaksimalkan profitabilitas. Semakin tinggi TP, semakin efektif strategi dalam menyeimbangkan peningkatan permintaan dan penurunan harga jual. Kedua, penelitian ini menggunakan *Profit Improvement Percentage* (PIP) untuk mengukur peningkatan kinerja relatif strategi Q-Learning dibandingkan dengan masing-masing skenario diskon tetap. PIP menunjukkan seberapa besar keuntungan tambahan yang diperoleh dengan menggunakan metode adaptif ini. Nilai positif berarti strategi Q-Learning menghasilkan profit yang lebih tinggi, sedangkan nilai negatif menunjukkan kinerja yang lebih rendah dibanding *baseline* diskon tetap.

3. HASIL DAN PEMBAHASAN

Hasil pada penelitian ini disusun sebagai berikut. Bagian 3.1 menjelaskan hasil prapemrosesan data dan distribusi *state*; Bagian 3.2 menganalisis konvergensi pelatihan agen Q-Learning; Bagian 3.3 menyajikan kebijakan diskon optimal dari Q-table yang telah konvergen; Bagian 3.4 membahas distribusi tingkat diskon yang diberikan agen;

Bagian 3.5 membandingkan total profit metode yang diusulkan terhadap lima *baseline* diskon tetap; dan Bagian 3.6 memberikan pembahasan lebih luas atas temuan penelitian.

3.1. Prapemrosesan Data dan Distribusi State

Tabel 5 merangkum statistik deskriptif variabel pendapatan dan frekuensi setelah tahap prapemrosesan. Variabel pendapatan menunjukkan distribusi yang sangat *right-skewed*, dengan nilai rata-rata sebesar 20,97 dan standar deviasi sebesar 291,12. Nilai pendapatan berkisar antara 0,001 hingga 168.469,60, dengan nilai maksimum yang jauh lebih tinggi dibandingkan nilai kuartil ketiga sebesar 17,96. Kondisi tersebut menunjukkan adanya sejumlah kecil transaksi dengan nilai yang sangat besar dibandingkan sebagian besar transaksi lainnya. Sementara itu, variabel frekuensi memiliki rata-rata sebesar 17,56 transaksi dengan standar deviasi 36,09. Frekuensi pembelian pelanggan berkisar antara 1 hingga 209 transaksi, dengan median sebesar 6 transaksi.

Tabel 5. Statistik Deskriptif *Revenue* dan Frekuensi Pembelian

Statistik	<i>R</i>	<i>Fr</i>
Jumlah	449.107	449.107
Rata-rata	20,97	17,56
Standar Deviasi	291,12	36,09
Minimum	0,001	1
Persentil ke-25 (Q_1)	4,25	3
Median	10,50	6
Persentil ke-75 (Q_3)	17,96	13
Maksimum	168.469,60	209
Jumlah	449.107	449.107
Rata-rata	20,97	17,56

Berdasarkan statistik tersebut, nilai pendapatan dikategorikan menjadi tiga tingkat menggunakan kuartil pertama ($Q_1 = 4,25$) dan kuartil ketiga ($Q_3 = 17,96$) sebagai ambang batas. Transaksi dengan pendapatan $\leq 4,25$ dikategorikan sebagai *Low Revenue*, transaksi dengan nilai antara 4,25 dan 17,96 dikategorikan sebagai *Medium Revenue*, sedangkan transaksi dengan nilai $> 17,96$ dikategorikan sebagai *High Revenue*. Sementara itu, frekuensi dibagi menjadi dua kategori berdasarkan nilai median, yaitu 6 transaksi. Pelanggan dengan frekuensi pembelian ≤ 6 transaksi dikategorikan sebagai *Low Frequency*, sedangkan pelanggan dengan frekuensi pembelian > 6 transaksi dikategorikan sebagai *High Frequency*. Kombinasi ketiga kategori pendapatan dua kategori frekuensi tersebut kemudian membentuk enam state yang digunakan dalam proses pembelajaran agen RL seperti yang terlihat pada Tabel 6.

Tabel 6. Distribusi State Pelanggan

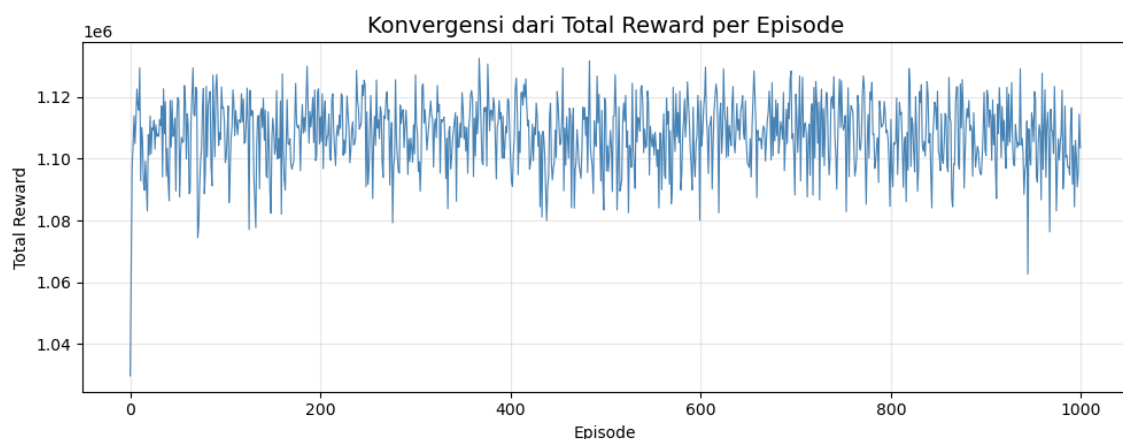
State	Jumlah Transaksi	Persentase (%)
Medium Revenue – Low Frequency	119.549	26.62%
Medium Revenue – High Frequency	106.321	23.67%
High Revenue – High Frequency	69.326	15.44%

Low Revenue – High Frequency	60.053	13.37%
Low Revenue – Low Frequency	50.623	11.27%
High Revenue – Low Frequency	43.235	9.63%

Pada Tabel 6 menunjukkan bahwa jumlah transaksi terbesar adalah *Medium Revenue–High Frequency*, yaitu sebanyak 119.549 transaksi, diikuti oleh *Medium Revenue–Low Frequency* sebanyak 106.321 transaksi. Sebaliknya, *state* dengan jumlah transaksi paling sedikit adalah *High Revenue–Low Frequency*, yaitu sebanyak 43.235 transaksi. Kondisi ini menunjukkan bahwa transaksi dengan tingkat pendapatan sedang mendominasi dataset, baik pada pelanggan dengan frekuensi pembelian rendah maupun tinggi. Sementara itu, transaksi dengan *High Revenue* relatif lebih sedikit, khususnya pada pelanggan dengan *Low Frequency*. Meskipun distribusi keenam *state* tidak sepenuhnya seimbang, seluruh kondisi pelanggan memiliki jumlah observasi yang memadai untuk digunakan dalam proses pembelajaran Q-Learning.

3.2. Konvergensi Pelatihan Q-Learning

Gambar 2 menyajikan total *reward* yang diperoleh agen Q-Learning selama 1.000 episode pelatihan. Total *reward* merepresentasikan profit kumulatif dari seluruh transaksi dalam satu episode berdasarkan formulasi *reward* yang disesuaikan elastisitas harga.



Gambar 2. Kurva Konvergensi Q-Learning

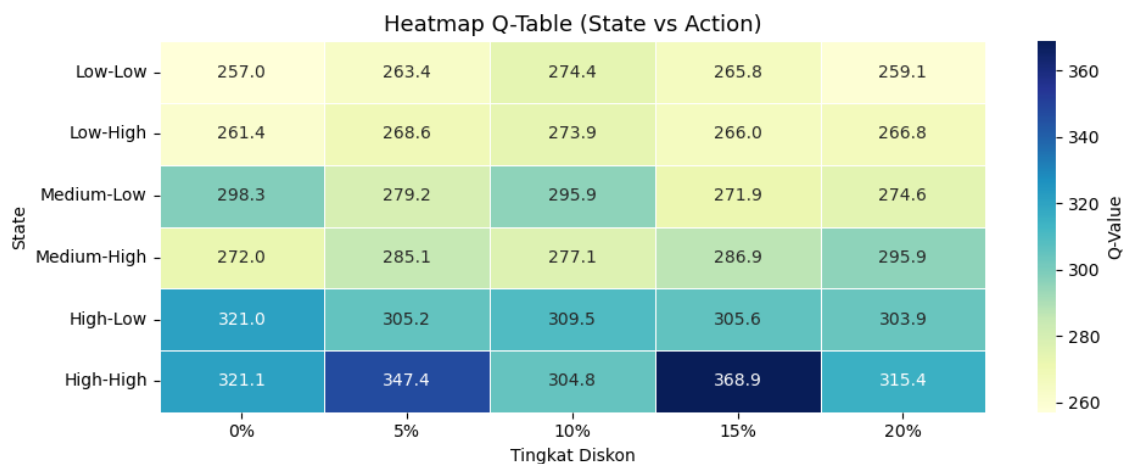
Berdasarkan Gambar 2, total *reward* menunjukkan fluktuasi cukup besar pada beberapa episode awal, kemudian relatif stabil pada kisaran $1,08 \times 10^6$ hingga $1,13 \times 10^6$. Tidak terlihat tren peningkatan atau penurunan yang konsisten hingga episode ke-1.000, yang menunjukkan bahwa kinerja agen telah mencapai kondisi yang relatif stabil. Fluktuasi yang masih terjadi dapat dipengaruhi oleh mekanisme eksplorasi ϵ -greedy dan variasi urutan transaksi selama pelatihan.

3.3. Kebijakan Diskon Optimal

Setelah pelatihan Q-Learning selesai, kebijakan diskon optimal diekstraksi dari Q-table yang telah konvergen dengan memilih aksi bernilai Q tertinggi untuk setiap *state*. Kebijakan yang dihasilkan dan nilai Q-table disajikan pada Tabel 7 dan Gambar 3.

Tabel 7. Kebijakan Diskon Optimal dari Q-table yang Telah Konvergen

State	Kategori <i>Revenue</i>	Kategori <i>Frequency</i>	Diskon Optimal
1	<i>Low</i>	<i>Low</i>	10%
2	<i>Low</i>	<i>High</i>	10%
3	<i>Medium</i>	<i>Low</i>	0%
4	<i>Medium</i>	<i>High</i>	20%
5	<i>High</i>	<i>Low</i>	0%
6	<i>High</i>	<i>High</i>	15%



Gambar 3. Heatmap Q-table

Berdasarkan hasil pada Tabel 6, kebijakan optimal yang dipelajari menunjukkan pola yang lebih beragam antar-state. Pada kategori frekuensi *Low*, agen memilih diskon 10% untuk state *Low Revenue–Low Frequency* dan *Low Revenue–High Frequency*, sementara pada state *Medium Revenue–Low Frequency* dan *High Revenue–Low Frequency*, agen justru memilih diskon 0%. Untuk kategori frekuensi *High*, agen memilih diskon 20% pada *Medium Revenue–High Frequency*, sedangkan pada *High Revenue–High Frequency* diskon optimal yang dipilih adalah 15%. Pola ini menunjukkan bahwa keputusan diskon tidak hanya ditentukan oleh frekuensi pembelian, tetapi juga mempertimbangkan tingkat pendapatan pelanggan. Dengan demikian, Q-Learning mampu membedakan tingkat insentif berdasarkan karakteristik state untuk memperoleh nilai *reward* kumulatif yang lebih tinggi, dibandingkan menerapkan satu tingkat diskon yang sama pada seluruh kelompok pelanggan.

Di sisi lain, nilai Q pada Gambar 3 menunjukkan kecenderungan meningkat pada state dengan tingkat pendapatan dan frekuensi yang lebih tinggi. Nilai Q maksimum sebesar 368,87 diperoleh pada state *High Revenue–High Frequency* dengan diskon 15%, sedangkan nilai Q terendah sebesar 257,04 terdapat pada state *Low Revenue–Low Frequency* dengan diskon 0%. Tingginya nilai Q pada pelanggan dengan karakteristik *High Revenue* dan *High Frequency* mengindikasikan bahwa state tersebut memiliki kontribusi *reward* kumulatif paling besar. Menariknya, pada state ini diskon

15% memberikan nilai Q yang lebih tinggi dibandingkan diskon 20%, sehingga diskon yang lebih besar tidak selalu menghasilkan *reward* yang lebih optimal.

3.4. Distribusi Diskon

Untuk memahami lebih lanjut perilaku kebijakan yang dipelajari, distribusi tingkat diskon yang diberikan agen Q-Learning pada seluruh transaksi yang dianalisis. Tabel 8 merangkum jumlah transaksi dan persentase untuk setiap tingkat diskon.

Tabel 8. Distribusi Tingkat Diskon

Tingkat Diskon	Jumlah Transaksi	Persentase (%)
0%	149.556	33,30
5%	0	0,00
10%	110.676	24,64
15%	69.326	15,44
20%	119.549	26,62

Berdasarkan Tabel 8, terlihat bahwa agen Q-Learning paling sering memilih diskon 0%, dengan jumlah 149.556 transaksi atau 33,30% dari seluruh transaksi. Pilihan berikutnya adalah diskon 20% sebesar 26,62%, diikuti oleh diskon 10% sebesar 24,64% dan diskon 15% sebesar 15,44%. Sementara itu, diskon 5% tidak pernah dipilih oleh agen. Distribusi tersebut menunjukkan bahwa agen tidak secara umum cenderung memberikan diskon tinggi, melainkan menyesuaikan tingkat diskon dengan karakteristik state pelanggan. Tingginya proporsi diskon 0% mengindikasikan bahwa pada sejumlah kondisi pelanggan, pemberian diskon tidak memberikan tambahan *reward* yang cukup besar untuk mengompensasi penurunan pendapatan akibat pengurangan harga.

3.5. Perbandingan Profit

Tabel 9 menyajikan total profit yang dihasilkan oleh setiap skenario eksperimen pada seluruh 397.885 transaksi. Strategi diskon dinamis Q-Learning yang diusulkan dibandingkan dengan lima *baseline* diskon tetap mulai dari 0% hingga 20%.

Table 9. Peningkatan Profit Q-Learning Relatif terhadap Setiap *Baseline* Diskon Tetap.

Skenario	Metode	Total Profit	Peningkatan (%)
1	Diskon tetap 0%	9.418.006,75	+13.46%
2	Diskon tetap 5%	8.947.106,42	+19.71%
3	Diskon tetap 10%	8.476.206,08	+25.96%
4	Diskon tetap 15%	8.005.305,74	+32.21%
5	Diskon tetap 20%	7.534.405,4	+38.46%
6	Q-Learning (diusulkan)	10.432.467,3	-

Seperti yang terlihat di Tabel 9, penerapan diskon tetap menunjukkan kecenderungan penurunan profit seiring dengan meningkatnya tingkat diskon. Sehingga, pemberian diskon yang lebih besar tidak menghasilkan peningkatan profit, karena pengurangan margin akibat diskon lebih besar daripada tambahan pendapatan yang diperoleh dari

peningkatan permintaan. Sebaliknya, metode Q-Learning menghasilkan total profit tertinggi sebesar 10.432.467,30. Nilai tersebut lebih tinggi dibandingkan seluruh skenario diskon tetap. Keunggulan Q-Learning juga semakin terlihat ketika dibandingkan dengan baseline diskon yang lebih tinggi.

Hasil tersebut menunjukkan bahwa Q-Learning mampu menghasilkan profit yang sedikit lebih tinggi dibandingkan strategi tanpa diskon, sekaligus mengungguli seluruh skenario diskon tetap. Keunggulan ini berasal dari kemampuan agen untuk memilih tingkat diskon secara adaptif berdasarkan kondisi pelanggan, sebagaimana ditunjukkan pada kebijakan optimal sebelumnya. Agen tidak menerapkan satu tingkat diskon secara seragam, tetapi memilih 0%, 10%, 15%, atau 20% sesuai dengan state pelanggan, sementara diskon 5% tidak pernah menjadi pilihan optimal. Strategi adaptif tersebut memungkinkan pemberian diskon hanya ketika diperkirakan memberikan kontribusi *reward* yang menguntungkan, sehingga margin dapat dipertahankan pada kondisi pelanggan yang tidak memerlukan insentif harga.

Dengan demikian, hasil pada Tabel 9 memperkuat bahwa pendekatan Q-Learning lebih efektif dibandingkan strategi diskon tetap dalam memaksimalkan profit pada penelitian ini. Menariknya, strategi Q-Learning tidak bergantung pada pemberian diskon maksimum, tetapi menghasilkan profit tertinggi melalui kombinasi tingkat diskon yang selektif. Bahkan, ketika dibandingkan dengan strategi tanpa diskon yang merupakan baseline dengan profit tertinggi di antara seluruh strategi tetap, Q-Learning masih mampu memberikan peningkatan profit sebesar 10,77%. Temuan ini menunjukkan bahwa keuntungan utama Q-Learning terletak pada kemampuannya mengoptimalkan keputusan diskon secara adaptif berdasarkan karakteristik pelanggan, sehingga trade-off antara peningkatan permintaan dan penurunan margin dapat dikelola dengan lebih efektif.

3.6. Pembahasan

Hasil dari penelitian ini menunjukkan bahwa strategi Q-Learning mampu mempelajari kebijakan diskon yang interpretatif dan efektif. Kebijakan yang dipelajari secara konsisten memberikan diskon lebih tinggi kepada pelanggan berfrekuensi tinggi, menunjukkan bahwa keterlibatan pelanggan lebih berperan dibandingkan nilai transaksi dalam menentukan diskon optimal, meskipun pendapatan memengaruhi besaran nilai Q, frekuensi pembelian tetap menjadi faktor utama pendorong alokasi diskon. Dari sisi profitabilitas, Q-Learning mampu menghasilkan profit sebesar 10.432.467,30, mengungguli semua *baseline* dan mampu meningkatkan profit, sehingga membuktikan alokasi diskon adaptif lebih efektif dibandingkan kebijakan statis.

Lebih lanjut, fungsi *reward* yang didasarkan pada elastisitas harga berperan penting dalam menentukan kebijakan diskon yang dihasilkan oleh agen. Dengan mempertimbangkan respons permintaan terhadap perubahan harga, fungsi *reward* mendorong agen untuk menyeimbangkan potensi peningkatan permintaan dengan penurunan margin akibat pemberian diskon. Hasil penelitian menunjukkan bahwa agen tidak selalu memilih diskon yang tinggi, tetapi menyesuaikan keputusan berdasarkan kondisi pelanggan, dengan tingkat diskon optimal yang mencakup 0%, 10%, 15%, dan 20%, sedangkan diskon 5% tidak menjadi pilihan optimal pada seluruh *state*. Dominasi pilihan diskon 0%, 10%, dan 20% pada distribusi transaksi juga menunjukkan bahwa agen mampu memberikan diskon secara selektif, bukan menerapkannya secara seragam.

Meskipun demikian, penelitian ini memiliki beberapa keterbatasan. Pertama, parameter elastisitas harga bersifat simulasi, bukan estimasi dari data permintaan aktual. Kedua, transisi *state* diperkirakan melalui sampling acak karena dataset tidak menyediakan respons pelanggan secara berurutan. Ketiga, representasi *state* disederhanakan hanya mencakup pendapatan dan frekuensi. Penelitian selanjutnya dapat mengatasi hal ini dengan elastisitas yang diestimasi secara empiris, memperkaya fitur perilaku pelanggan, serta menerapkan arsitektur *Deep Q-Network* (DQN) untuk menangani ruang keadaan yang lebih kompleks dan lingkungan yang bersifat *real-time*.

4. KESIMPULAN

Penelitian ini mengembangkan model diskon dinamis pada platform e-commerce menggunakan algoritma Q-Learning dengan tujuan memaksimalkan keuntungan melalui pemberian diskon yang disesuaikan dengan kondisi pelanggan. Agen memilih salah satu dari lima tingkat diskon (0–20%), kemudian mempelajari kebijakan optimal berdasarkan karakteristik pendapatan dan frekuensi pembelian pelanggan. Hasil penelitian menunjukkan bahwa kebijakan yang diperoleh bersifat adaptif, dengan pilihan diskon optimal yang bervariasi antara 0%, 10%, 15%, dan 20% pada masing-masing *state*, sementara diskon 5% tidak menjadi pilihan optimal. Secara keseluruhan, Q-Learning menghasilkan total profit tertinggi dibandingkan seluruh strategi diskon tetap. Temuan ini menunjukkan bahwa strategi diskon berbasis Q-Learning mampu memberikan keputusan yang lebih adaptif dan efektif dalam memaksimalkan profit dibandingkan penerapan satu tingkat diskon secara tetap. Keberhasilan model juga didukung oleh integrasi elastisitas harga ke dalam fungsi *reward*, sehingga agen dapat mempertimbangkan respons permintaan terhadap perubahan harga dalam menentukan tingkat diskon yang optimal. Untuk penelitian selanjutnya, estimasi elastisitas harga dapat dilakukan secara empiris menggunakan data transaksi nyata, serta model dapat dikembangkan menggunakan arsitektur Deep Q-Network (DQN) agar mampu menangani fitur pelanggan yang lebih kompleks, seperti kebaruan pembelian, kategori produk, dan faktor musiman. Selain itu, penerapan model pada lingkungan e-commerce nyata dengan mekanisme umpan balik pelanggan secara langsung dapat menjadi langkah lanjutan untuk menguji efektivitas dan meningkatkan kesiapan model diskon berbasis Reinforcement Learning dalam konteks industri ritel.

DAFTAR PUSTAKA

- [1] A. N. Raji, A. O. Olawore, and D. Osahor, “The impact of E-commerce giants on SMEs: Challenges, opportunities, and the fight for survival in the digital economy,” *World J. Adv. Res. Rev.*, vol. 20, no. 2, pp. 1412–1433, 2023.
- [2] A. Julian, C. Maheswar, R. Ramyadevi, M. S. M. Dhar, and S. Selvi, “Analysis and estimation of discounts on product prices in e-commerce sites,” in *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, 2024, pp. 759–762.
- [3] V. Pabalkar, “Price—impact on consumer behavior through e-commerce and future opportunities,” in *Pandemic to Endemic*, Routledge, 2024, pp. 222–239.
- [4] F. Cherbonnier and C. Gollier, “Fixing our public discounting systems,” *Annu. Rev. Financ. Econ.*, vol. 15, no. 1, pp. 147–164, 2023.
- [5] H. Li and Zh. Chen, “How do e-commerce platforms and retailers implement

-
- discount pricing policies under consumers are strategic?,” *PLoS One*, vol. 19, no. 5, p. e0296654, 2024.
- [6] S. Sharma, P. Singh, A. Malik, A. Kaur, and C. Irawan, “Comprehensive Literature Review on Machine Learning--Driven Dynamic Pricing Strategies in E-Commerce,” in *2025 International Conference on Metaverse and Current Trends in Computing (ICMCTC)*, 2025, pp. 1–9.
 - [7] A. H. Lubis, P. Sihombing, E. B. Nababan, and R. F. Rahmat, “Reformulating Q-Learning as a Hybrid Classifier for Predictive Marketing Analytics,” *Int. J. Intell. Eng. Syst.*, vol. 18, no. 11, pp. 29–44, 2025.
 - [8] A. K. Shakya, G. Pillai, and S. Chakrabarty, “Reinforcement learning algorithms: A brief survey,” *Expert Syst. Appl.*, vol. 231, p. 120495, 2023.
 - [9] G. Vivar-Estudillo, M. Estudillo-Ayala, J. Beltrán-Hernández, M. Ibarra-Manzano, and C. Lastre-Domínguez, “Industrial Applications of Q-Learning: A Systematic Review,” *WSEAS Trans. Inf. Sci. Appl.*, vol. 22, pp. 525–539, 2025.
 - [10] A. H. Lubis, E. B. Nababan, P. Sihombing, and R. F. Rahmat, “Comparison of Q-Learning and SARSA in Allocating Marketing Budget,” in *2025 International Conference on Artificial Intelligence and Technological Solutions (ICAITech)*, 2025, pp. 14–20.
 - [11] M. Apte, P. Datar, K. Kale, and P. R. Deshmukh, “Dynamic retail pricing via q-learning-a reinforcement learning framework for enhanced revenue management,” in *2025 1st International Conference on AIML-Applications for Engineering & Technology (ICAET)*, 2025, pp. 1–5.
 - [12] P. Mullapudi, “A Reinforcement Learning Approach to Dynamic Pricing,” *IJSAT-International J. Sci. Technol.*, vol. 16, no. 4, 2025.
 - [13] H. MuhamedAle *et al.*, “Dynamic Pricing Algorithm Using Reinforcement Learning in Online Retail Platforms,” in *2025 3rd International Conference on Cyber Resilience (ICCR)*, 2025, pp. 1–7.
 - [14] A. Amato and V. Di Lecce, “Data preprocessing impact on machine learning algorithm performance,” *Open Comput. Sci.*, vol. 13, no. 1, p. 20220278, 2023.
 - [15] Q. W. Khan, “Exploring Markov decision processes: a comprehensive survey of optimization applications and techniques,” *Igmin Res.*, vol. 2, no. 7, pp. 508–517, 2024.
 - [16] X. Liu, “Dynamic coupon targeting using batch deep reinforcement learning: An application to livestream shopping,” *Mark. Sci.*, vol. 42, no. 4, pp. 637–658, 2023.
 - [17] S. Mehta and S. S. Sarpal, “Maximizing Privacy in Reinforcement Learning with Federated Approaches,” in *2024 4th International Conference on Intelligent Technologies (CONIT)*, 2024, pp. 1–5.
 - [18] T. Tan, H. Xie, and D. Lian, “Adaptive order Q-learning,” in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024, pp. 4946–4954.
 - [19] Z. Zhao, “Variants of Bellman equation on reinforcement learning problems,” in *2nd International Conference on Artificial Intelligence, Automation, and High-Performance Computing (AIAHPC 2022)*, 2022, pp. 470–481.
 - [20] M. Hu, “Temporal difference learning,” in *The Art of Reinforcement Learning: Fundamentals, Mathematics, and Implementations with Python*, Springer, 2023, pp. 75–107.