VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8286

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

REINFORCEMENT LEARNING-BASED DYNAMIC PRICING IN A STOCHASTIC DEMAND–SUPPLY ENVIRONMENT

Nur Alamsyah[1*]; Budiman[1]; Almira Nurchawilah[1]; Wala Erpurini[2]

Information System[1]
Universitas Informatika Dan Bisnis Indonesia, Bandung, Indonesia[1]
http://www.unibi.ac.id[1]
nuralamsyah@unibi.ac.id*, budiman@unibi.ac.id, almira@unibi.ac.id

Management[2]
Universitas Jenderal Ahmad Yani, Cimahi, Indonesia[2]
http://www.unjani.ac.id[2]
walaerpurini@mn.unjani.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Dynamic pricing in ride-sharing platforms must balance revenue generation with stable pricing decisions under changing demand and supply. This study aims to develop and evaluate a reinforcement learning-based dynamic pricing policy that maximizes expected revenue while reducing abrupt policy-level price adjustments. A stochastic contextual environment was constructed from 1,000 historical ride records and evaluated using a leakage-safe 70/15/15 train-validation-test split. The agent was trained with Proximal Policy Optimization (PPO) using five discrete price adjustments from -10% to +10%. Expected revenue was combined with a multiplier-based stability penalty, where stability was measured from changes in the price multiplier rather than nominal price variation across heterogeneous rides. Across 30 paired test episodes, the PPO policy achieved a cumulative reward of 103,316.25 +/- 3,243.99 and expected revenue of 103,449.18 +/- 3,241.27, significantly exceeding static pricing ($p < 0.001$). Relative to rule-based surge pricing, PPO produced statistically indistinguishable cumulative reward ($p = 0.808$) while reducing multiplier volatility by 24.83%, mean absolute multiplier change by 24.21%, and action switch rate by 10.97% (all $p < 0.001$). These results indicate that PPO can preserve near-surge revenue while producing smoother dynamic pricing decisions within the simulated environment.*

Keywords: *Dynamic Pricing, Multiplier Stability, Proximal Policy Optimization, Reinforcement Learning, Revenue Optimization.*

Intisari— *Penetapan harga dinamis pada layanan transportasi daring perlu menyeimbangkan pendapatan dengan kestabilan keputusan harga ketika permintaan dan penawaran berubah. Penelitian ini bertujuan mengembangkan dan mengevaluasi kebijakan harga dinamis berbasis reinforcement learning yang memaksimalkan pendapatan harapan sekaligus mengurangi perubahan kebijakan harga yang mendadak. Lingkungan kontekstual stokastik dibangun dari 1.000 data historis perjalanan dan dievaluasi menggunakan pembagian data 70/15/15 untuk pelatihan, validasi, dan pengujian tanpa kebocoran data. Agen dilatih menggunakan Proximal Policy Optimization (PPO) dengan lima aksi penyesuaian harga diskret dari -10% hingga +10%. Pendapatan harapan digabungkan dengan penalti kestabilan berbasis multiplier harga, sehingga kestabilan diukur dari perubahan multiplier, bukan dari variasi harga nominal antarperjalanan yang heterogen. Pada 30 episode uji berpasangan, kebijakan PPO memperoleh cumulative reward sebesar 103.316,25 +/- 3.243,99 dan expected revenue sebesar 103.449,18 +/- 3.241,27, yang secara signifikan lebih tinggi daripada harga statis ($p < 0,001$). Dibandingkan surge pricing berbasis aturan, PPO menghasilkan cumulative reward yang tidak berbeda signifikan ($p = 0,808$), tetapi menurunkan volatilitas multiplier sebesar*

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8286

24,83%, perubahan multiplier absolut sebesar 24,21%, dan action switch rate sebesar 10,97% (seluruhnya p < 0,001). Hasil ini menunjukkan bahwa PPO dapat mempertahankan pendapatan yang mendekati surge pricing dengan keputusan harga yang lebih halus dalam lingkungan simulasi.

***Kata Kunci**: Penetapan Harga Dinamis, Stabilitas Pengali, Optimisasi Kebijakan Proksimal, Pembelajaran Penguatan, Optimasi Pendapatan.*

INTRODUCTION

Dynamic pricing has been widely adopted in service-based platforms as a mechanism to adjust prices in response to changing market conditions [1], [2]. In industries such as ride-sharing, transportation, and on-demand services, pricing decisions are influenced by fluctuating demand, varying supply availability, temporal factors, and heterogeneous customer behavior [3], [4]. Conventional pricing strategies commonly rely on static pricing schemes or simple rule-based surge mechanisms, where prices are adjusted using predefined thresholds derived from demand–supply ratios [5], [6]. While such approaches are straightforward and computationally efficient, they often lack the flexibility required to respond effectively to complex and rapidly changing market environments [7], [8].

A substantial body of prior research has investigated dynamic pricing using analytical and data-driven approaches [9], [10], [11]. Early studies primarily focused on rule-based and optimization-based pricing models, where prices were determined using predefined demand functions or economic equilibrium assumptions [12]. These models provided valuable theoretical insights but often relied on strong assumptions regarding customer behavior and market stability [13], [14]. As a result, their applicability in real-world settings with high uncertainty and non-stationary demand has been limited. In practice, pricing decisions are frequently influenced by factors that are difficult to model explicitly, such as sudden demand surges, uneven supply distribution, and varying customer sensitivity to price changes.

More recent studies have explored machine learning-based approaches to dynamic pricing, particularly using supervised learning techniques [15], [16], [17]. These methods typically predict optimal prices based on historical data by learning relationships between contextual features and observed pricing outcomes. While supervised models have shown improved predictive performance compared to traditional analytical models, they generally treat pricing as a static prediction problem [18]. Consequently, they do not explicitly account for the sequential and interactive nature of pricing decisions, where current pricing actions can influence future demand, supply availability, and customer behavior [19], [20]. This limitation restricts their ability to adapt pricing policies dynamically in response to evolving market conditions.

In response to these challenges, reinforcement learning has emerged as a promising paradigm for dynamic pricing problems characterized by uncertainty and sequential decision-making [21]. Unlike supervised learning, reinforcement learning enables an agent to learn pricing policies through direct interaction with an environment, using feedback signals to guide policy improvement over time [22], [23], [24]. Prior studies have demonstrated the potential of reinforcement learning for pricing and revenue management tasks [25], [26]; however, many existing works focus primarily on revenue maximization and often overlook the issue of price stability. In practical applications, excessive price fluctuations may lead to customer dissatisfaction and reduced long-term trust, highlighting the importance of considering stability alongside profitability.

This study addressed the aforementioned gap by formulating the dynamic pricing problem as a reinforcement learning task under demand–supply uncertainty, with explicit consideration of both revenue optimization and price stability. A simulation-based environment was developed using historical ride-sharing data to capture realistic variations in market conditions, including demand intensity, supply availability, temporal factors, and customer characteristics. A reinforcement learning agent based on Proximal Policy Optimization was employed to learn adaptive pricing policies through interaction with this environment.

The proposed approach was evaluated against two commonly used baseline strategies, namely static pricing and rule-based surge pricing, to provide a comprehensive comparison of pricing performance. The evaluation focused on average revenue and price volatility as key performance indicators, thereby enabling an explicit analysis of the trade-off between profitability and stability. The results demonstrated that while rule-based surge pricing achieved the highest revenue, it also produced the largest price fluctuations. In contrast,

the reinforcement learning-based approach achieved a more balanced performance, delivering competitive revenue gains with lower price volatility. The proposed framework was evaluated against static pricing and rule-based surge pricing under identical paired test scenarios. Economic performance was assessed using cumulative reward and cumulative expected revenue, whereas policy stability was assessed using multiplier volatility, mean absolute multiplier change, and action switch rate.

Unlike prior reinforcement learning pricing studies that primarily optimize revenue or assess stability through nominal price variation, this study does not claim to introduce a new reinforcement learning algorithm. Instead, it develops a reproducible stability-aware dynamic pricing framework based on Proximal Policy Optimization (PPO) for ride-sharing contexts with uncertain demand and supply conditions. At each decision step, PPO selects a discrete relative price adjustment $a_t \in \{-0.10, -0.05, 0, 0.05, 0.10\}$, which determines the offered price as $p_t = b_t(1 + a_t) = b_t m_t$. The reward is defined as $r_t = ER_t - \lambda s_b |m_t - m_{(t-1)}|$, where $ER_t = p_t P(accept_t)$, s_b is the training-derived base-price scale, and λ is selected through validation. Thus, policy stability is measured from changes in the price multiplier rather than nominal price differences that may be confounded by heterogeneous base fares.

Accordingly, this study makes three contributions. First, it formulates ride-sharing dynamic pricing as a sequential decision-making problem using PPO with a bounded and operationally interpretable discrete multiplier action space. Second, it integrates multiplier-based reward shaping and policy-stability measures, including multiplier volatility, mean absolute multiplier change, and action switch rate, to evaluate revenue–stability trade-offs without conflating policy adjustments with ride-specific base-price variation.

Third, it implements a leakage-safe train-validation-test protocol and evaluates PPO against static pricing and rule-based surge pricing using identical paired test scenarios, followed by Friedman tests and Holm-adjusted Wilcoxon signed-rank comparisons. Collectively, these contributions provide statistically supported evidence on how a PPO-based policy can preserve near-surge cumulative reward while producing smoother pricing-policy adjustments within the simulated ride-sharing environment.

MATERIALS AND METHODS

This study employed a simulation-based reinforcement learning framework to develop and evaluate a dynamic pricing policy under stochastic demand–supply conditions. Figure 1 summarizes the workflow, including leakage-safe data preparation, state construction, PPO training and validation, paired baseline comparison, and statistical evaluation. The 1,000 ride-sharing records were split into training, validation, and test subsets using a 70/15/15 ratio before preprocessing. The preprocessing transformer was fitted only on the training data. The final state consisted of 18 encoded market-context features, the normalized ride-specific base price, and the previous price multiplier.

The dynamic pricing environment was formulated as a contextual Markov decision process in which PPO selected one of five discrete price adjustments, {−0.10, −0.05, 0, 0.05, 0.10}, at each step. The reward combined expected revenue with a multiplier-based stability penalty, $r_t = ER_t - \lambda s_b |m_t - m_{(t-1)}|$, where λ was selected using the validation set. The resulting policy was evaluated against static and rule-based surge pricing using 30 paired test episodes of 300 steps. Performance was assessed through cumulative reward, cumulative expected revenue, multiplier volatility, mean absolute multiplier change, and action switch rate, followed by Friedman and Holm-adjusted Wilcoxon signed-rank test

Data Collection

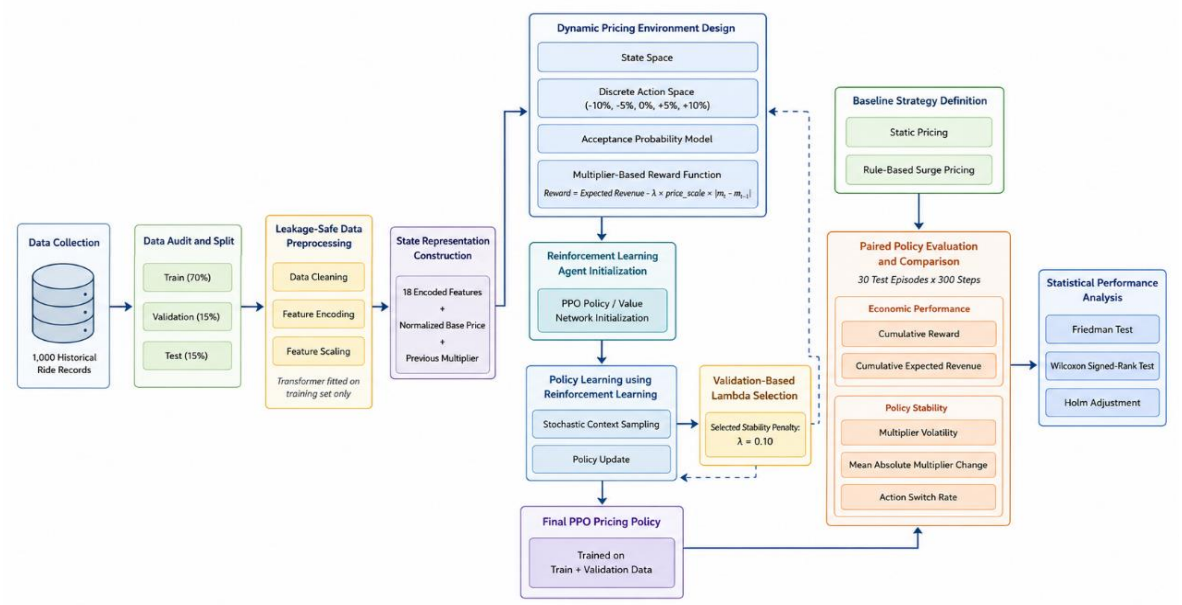
The dataset used in this study was obtained from the Kaggle data science platform, (https://www.kaggle.com/datasets/arashnic/dynamic-pricing-dataset) which provides publicly accessible datasets for research and educational purposes. The selected dataset contains historical ride-sharing transaction records and was chosen because it reflects realistic demand–supply dynamics that are relevant to the dynamic pricing problem addressed in this research.

The dataset contains 1,000 ride-sharing records with variables describing rider demand, driver availability, location category, customer loyalty status, past ride history, average ratings, booking time, vehicle type, expected ride duration, and historical ride cost. These variables were selected because they provide observable demand-side, supply-side, temporal, and customer-context information for constructing a contextual dynamic pricing environment. An initial data audit found no duplicated records and no missing values in the variables used for the experiment.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8286

Historical_Cost_of_Ride was retained as the ride-specific reference base price rather than treated as a prediction target. It was used to anchor the price generated by PPO, static pricing, and rule-based surge pricing. To prevent information leakage, the records were split before preprocessing into 700 training, 150 validation, and 150 held-out test observations using seed 2026. Feature scaling and categorical encoding were fitted only on the training subset and then applied unchanged to the validation and test subsets. The validation subset was used to select the stability-penalty coefficient, whereas the test subset was reserved for the final paired policy comparison. Although the dataset is limited to 1,000 public records and does not represent a live operational stream, it provides a transparent proof-of-concept context for evaluating the revenue–stability trade-off under controlled simulated demand–supply conditions.



Source: (Research results, 2026)

Figure 1. Leakage-Safe Multiplier-Stability Dynamic Pricing Framework

Data Audit and Experimental Split

Before model development, the dataset was audited to identify duplicated observations, missing values, and inconsistencies in the variables used for the experiment. The audit confirmed that the selected variables contained no duplicated records and no missing values.

To prevent information leakage, the dataset was divided into training, validation, and test subsets before any preprocessing procedure was performed. The split followed a 70/15/15 ratio, resulting in 700 training observations, 150 validation observations, and 150 test observations. The training subset was used to learn the PPO policy, whereas the validation subset was used to select the stability-penalty coefficient. The held-out test subset was reserved exclusively for the final paired comparison among PPO, static pricing, and rule-based surge pricing.

All experiments used a fixed random seed of 2026 to support reproducibility. This split design ensured that information from the validation and test subsets did not influence feature transformation, penalty selection, or final performance evaluation.

Leakage-Safe Data Preprocessing

A leakage-safe preprocessing procedure was applied to ensure that information from the validation and test subsets did not influence model development. All preprocessing steps were fitted exclusively on the training subset and were then applied unchanged to the validation and test subsets.

The preprocessing stage included data cleaning, categorical feature encoding, and numerical feature scaling. Numerical variables, including the number of riders, number of drivers, number of past rides, average ratings, and expected ride duration, were standardized using parameters estimated from the training data. Categorical variables, including location category, customer loyalty status, time of booking, and vehicle type, were converted into numerical representations

using one-hot encoding with unknown-category handling.

The resulting transformation generated 18 encoded market-context features. *Historical_Cost_of_Ride* was intentionally excluded from this encoded feature vector because it was retained separately as the ride-specific base-price anchor. This separation ensured that the reference fare was not directly embedded in the market-context representation used by the policy network.

In addition, the training median of the historical ride cost was stored as a fixed base-price scale for state normalization and multiplier-based reward calculation. The fitted preprocessing transformer, data split manifest, and random seed were retained to support reproducibility of the experimental procedure.

State Representation Construction

State representation determines the information available to the reinforcement learning agent when selecting a pricing action. The preprocessing pipeline first produces an encoded market-context vector, whereas the complete pricing state additionally includes the normalized ride-specific base price and the multiplier selected at the preceding decision step.

$$\mathbf{x}_t = [x_{t,1}, x_{t,2}, \dots, x_{t,18}] \quad (1)$$

Here, x_t denotes the 18 encoded market-context features obtained after feature scaling and one-hot encoding. These features capture rider demand, driver availability, customer characteristics, temporal context, and ride-specific information.

The contextual vector x_t preserves the operational information available at each decision step without introducing handcrafted indicators. The complete state used by PPO is specified in the following State Space subsection.

Although the encoded representation forms a continuous feature space, its dimensionality remains manageable for the PPO model. After preprocessing, the state input contains 18 encoded market-context features. Dimensionality reduction was therefore not applied because removing features could reduce the interpretability of operational factors that may influence pricing decisions. Instead, potential sample-complexity issues were mitigated through standardized input features, a compact PPO policy and value network with two hidden layers of 64 units each, entropy regularization with a coefficient of 0.01, and a fixed train-validation-test protocol.

For the final evaluation, PPO, static pricing, and rule-based surge pricing were exposed to the same exogenous test scenarios. This paired evaluation design ensured that performance differences were attributable to the pricing policies rather than to differences in demand-supply contexts. The state representation therefore provided a consistent and reproducible basis for learning and evaluating dynamic pricing decisions.

By explicitly constructing the state representation from preprocessed real-world features, the proposed framework preserved the realism of the pricing environment while enabling the reinforcement learning agent to capture demand–supply dynamics in a data-driven manner. This state design served as the foundation for defining the dynamic pricing environment, which is described in the subsequent subsection.

State Space

The state space combines the processed ride context, the normalized ride-specific base price, and the multiplier selected at the preceding decision step:

$$s_t = \left[x_t, \frac{b_t}{s_b}, m_{t-1}\right. \quad (2)$$

Where (x_t) represents the 18 encoded market-context features, (b_t) is the historical ride cost used as the ride-specific base-price anchor, s_b is the training-derived base-price scale, and m_{t-1} is the multiplier selected at the previous decision step.

The resulting state contains 20 components. It incorporates demand-side, supply-side, customer-related, temporal, and ride-specific information while allowing the PPO agent to consider its most recent pricing decision. Including the previous multiplier also supports the evaluation of policy-level stability through multiplier changes rather than nominal-price differences across heterogeneous rides.

Action Space

The action space represented relative price adjustments selected by the PPO agent at each decision step. Rather than predicting an unrestricted continuous value, the agent selected one action from a bounded discrete action set:

$$\mathcal{A} = -0.10, -0.05, 0.00, 0.05, 0.10 \quad (3)$$

Let (a_t) denote the selected action at time step(t). The corresponding price multiplier is defined as:

$$m_t = 1 + a_t \quad (4)$$

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8286

The offered price is then calculated as:

$$p_t = b_t(1 + a_t) = b_t m_t \quad (5)$$

Where b_t denotes the ride-specific base price. This discrete formulation allows the policy to apply interpretable price adjustments, including 10% and 5% discounts, no adjustment, and 5% and 10% markups. It also ensures that PPO, static pricing, and rule-based surge pricing are evaluated using the same bounded pricing range.

Dynamic Pricing Environment Design

The dynamic pricing environment was formulated as a contextual Markov decision process that integrates the state representation, discrete action space, acceptance probability model, and multiplier-based reward function. At each decision step, the agent observes the current ride context, selects a relative price adjustment, and receives reward feedback.

Customer acceptance was modeled using a bounded sigmoid function that increases with demand-supply pressure, loyalty, ratings, and past ride history, and decreases as the offered price rises relative to the base price.:

$$P(accept_t) = \sigma[0.9(R_t/D_t - 1) - 2.0(p_t/b_t - 1) + 0.7L_t + 0.15(q_t - 4) + 0.002(h_t - 50) + 0.2 \quad (6)$$

Where σ(·) denotes the sigmoid function, R_t and D_t are the numbers of riders and drivers, L_t represents customer loyalty, q_t represents average ratings, and h_t denotes the number of past rides. Expected revenue is calculated as $ER_t = p_t \times P(accept_t)$

$$r_t = ER_t - \lambda s_b |m_t - m_{(t-1)} \quad (7)$$

Where m_t and $m_{(t-1)}$ are the current and previous price multipliers, s_b is the training-derived base-price scale, and λ controls the trade-off between expected revenue and multiplier stability. This multiplier-based penalty discourages abrupt policy adjustments without treating natural nominal-price differences across heterogeneous rides as instability.

During training, ride contexts were randomly sampled from the training pool, while PPO sampled actions from its policy distribution. These two mechanisms introduced stochasticity into the pricing environment by exposing the agent to varying demand-supply conditions and alternative pricing decisions. For the final evaluation, fixed test scenarios were applied consistently to PPO, static pricing, and rule-based surge pricing to ensure a paired and fair comparison.

The dynamic pricing environment integrates the state representation, discrete action space, acceptance probability model, and multiplier-based reward function. This structure allows the PPO agent to learn pricing decisions from reward feedback while considering both expected revenue and policy-level pricing stability.

PPO Training and Validation

The pricing policy was trained using Proximal Policy Optimization (PPO) implemented with the Stable-Baselines3 MlpPolicy. Both the policy and value networks used two hidden layers with 64 units per layer. Training used four parallel environments, n_steps = 1024, batch_size = 256, discount factor γ = 0.99, GAE coefficient = 0.95, clip range = 0.20, learning rate = 3×10^{-4}, and entropy coefficient = 0.01.

At each training step, PPO observed the state s_t, sampled an action a_t from its policy distribution, received reward r_t, and updated the policy using trajectories collected from the stochastic contextual environment. The training objective maximized the expected discounted return over an episode:

$$J(\theta) = E_{(\pi_\theta)}\left[\Sigma_{(t=0)}^{T} \gamma^t r_t\right] \quad (8)$$

PPO optimized a clipped surrogate objective to constrain excessively large policy updates and improve learning stability:

$$L^C LIP(\theta) = E_t\left[min\left(\rho_t(\theta)\hat{A}_t, clip(\rho_t(\theta), 1 - \varepsilon, 1 + \varepsilon)\hat{A}_t\right)\right] \quad (9)$$

.

Validation-Based Lambda Selection

The stability-penalty coefficient was selected on the validation subset from Λ = {0.00, 0.05, 0.10, 0.25, 0.50, 1.00}. For each candidate, PPO was trained for 30,000 time steps and evaluated on 30 paired validation episodes. Candidates were required to retain at least 95% of the highest validation expected revenue; among the eligible candidates, λ = 0.10 achieved the lowest multiplier volatility and was selected.

The final PPO policy was trained on the combined training and validation data for 100,000 time steps using random seed 2026. Stochastic action sampling was used during learning, whereas

deterministic actions were used during test evaluation. The learned PPO policy was then evaluated only on the held-out test scenarios and compared with the baseline policies described below.

Baseline Strategy Definition

Two baseline strategies were evaluated under the same exogenous test scenarios and acceptance model: static pricing and rule-based surge pricing.

Static Pricing

Static pricing retained the historical base price without adjustment. It represents a non-adaptive tariff and provides a reference for the value added by dynamic decisions.

$$m_t^s tatic = 1, p_t^s tatic = b_t \quad (10)$$

Because its multiplier remains constant, static pricing has zero multiplier volatility, zero multiplier change, and zero action switch rate by construction.

Rule-Based Surge Pricing

Rule-based surge pricing used the demand-supply ratio ρ_t = R_t/D_t to select a predefined relative adjustment. The strategy applied a −5% discount when ρ_t < 0.90, no adjustment when 0.90 ≤ ρ_t < 1.20, a 5% markup when 1.20 ≤ ρ_t < 1.50, and a 10% markup when ρ_t ≥ 1.50.

$$a_t^s urge \in -0.05, 0.00, 0.05, 0.10 according to \rho_t thresholds \quad (11)$$

The corresponding multiplier is $m_t^s urge = 1 + a_t^s urge$ and the offered price is $p_t^s urge = b_t m_t^s urge$. This baseline responds transparently to demand-supply imbalance but does not learn from revenue or stability outcomes. All policies were assessed on the same 30 fixed test scenarios. This paired protocol ensured that performance differences reflected policy decisions rather than different sequences of ride contexts.

Paired Policy Evaluation and Comparison

The final evaluation compared the PPO policy with static pricing and rule-based surge pricing under identical test conditions. All policies were evaluated using the same 30 paired test episodes, with each episode consisting of 300 pricing decisions. This paired design ensured that differences in performance were attributable to the pricing policies rather than to variations in demand-supply contexts or ride characteristics.

Economic performance was assessed using cumulative reward and cumulative expected revenue. Cumulative reward represents the total return obtained by each policy after accounting for the stability penalty, whereas cumulative expected revenue reflects the revenue outcome before the penalty is applied. Reporting both measures makes it possible to distinguish revenue generation from the effect of stability-oriented reward shaping.

Policy stability was evaluated using multiplier volatility, mean absolute multiplier change, and action switch rate. Multiplier volatility measures the overall variation of pricing multipliers within an episode, while mean absolute multiplier change captures the average magnitude of adjustment between consecutive decisions. Action switch rate indicates how frequently the policy changes its selected pricing action over time. Lower values of these metrics indicate smoother and more consistent pricing behaviour. These multiplier-based measures were used as the primary stability indicators because they reflect policy-induced adjustments without being confounded by differences in ride-specific base prices.

Performance Analysis

The final analysis compared PPO, static pricing, and rule-based surge pricing using the same 30 paired test episodes. Results were reported as mean and standard deviation across episodes. Economic performance was assessed using cumulative reward and cumulative expected revenue, whereas policy stability was assessed using multiplier volatility, mean absolute multiplier change, and action switch rate.

Global differences among the three policies were examined using the Friedman test. When a significant result was obtained, paired two-sided Wilcoxon signed-rank tests with Holm adjustment were used to identify significant pairwise differences. This procedure ensured that the reported revenue-stability trade-offs were evaluated using identical market scenarios and were not attributable to episode-level variation.

RESULTS AND DISCUSSION

This section presents the experimental results obtained from the proposed reinforcement learning–based dynamic pricing framework and discusses the observed findings. The results were generated through systematic evaluation across multiple episodes and were designed to reflect both

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

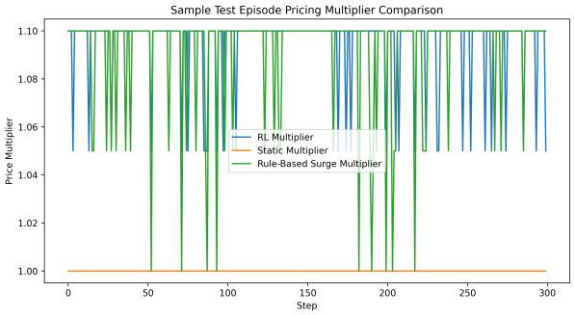
VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8286

revenue performance and pricing stability under varying demand–supply conditions.

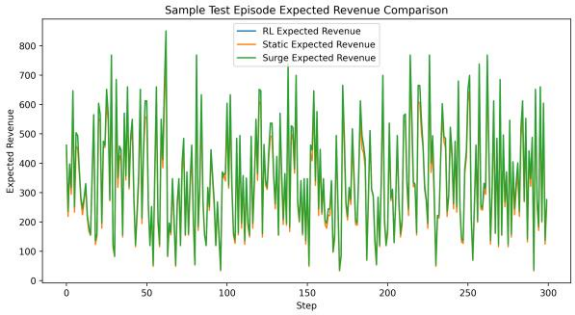
The analysis focused on comparing the learned pricing policy with baseline strategies, namely static pricing and rule-based surge pricing, using the evaluation metrics defined in the previous section. By employing consistent experimental settings and identical data inputs, the results offer a fair assessment of the effectiveness of each pricing approach.

The first set of experiments evaluated the performance of the proposed reinforcement learning based pricing policy in comparison with the baseline pricing strategy. The evaluation focused on the average cumulative reward obtained across multiple episodes, which represents the overall expected revenue generated by each pricing approach.

The experimental results show that the reinforcement learning policy consistently outperformed the baseline strategy in terms of average cumulative reward. The proposed method achieved an average reward of 96,578.10 with a standard deviation of 1,867.57, whereas the baseline pricing strategy yielded a lower average reward of 92,051.40 with a higher standard deviation of 3,272.93. These results indicate that the reinforcement learning agent was able to generate higher and more stable revenue outcomes compared to the non-learning baseline approach.



(a)



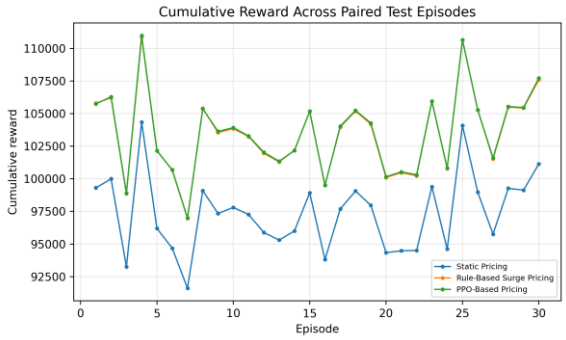
(b)

Source: (Research results, 2026)

Figure 2. Example multiplier and expected-revenue trajectories

.

(a)



(b)

Source: (Research results, 2026)

Figure 3. Cumulative Reward And Action Switch Rate Across Paired Test Episodes.

Figure 2 presents an illustrative paired test episode showing the multiplier trajectory and expected-revenue pattern produced by the evaluated pricing policies. The figure is intended to demonstrate the temporal behaviour of the policies under the same market context. However, the final conclusions of this study are based on the aggregated results from all 30 paired test episodes rather than on a single illustrative trajectory.

Figure 3(a) illustrates the price dynamics produced by the three strategies during a representative episode. The rule-based surge pricing approach exhibits more frequent and sharper price increases, particularly during high-demand periods, whereas the reinforcement learning policy demonstrates smoother and more adaptive price adjustments relative to the static baseline.

The corresponding expected revenue trajectories are presented in Figure 3(b). The surge-based strategy consistently generates higher revenue peaks, reflecting its aggressive pricing

behavior. In contrast, the reinforcement learning policy achieves comparable revenue levels while maintaining more moderate fluctuations, indicating a learned balance between revenue maximization and price stability.

Finally, Figure 3(c) shows the acceptance probability generated by the simulator. While rule-based surge pricing attains higher revenue, it also exhibits greater variability in acceptance probability, suggesting a potential trade-off between aggressive pricing and customer acceptance. Conversely, the reinforcement learning policy maintains a more stable acceptance probability over time, highlighting its ability to adapt pricing decisions in a way that preserves demand while still improving overall revenue.

Table 1. Multiplier-Based Policy-Stability Metrics Across Pricing Strategies

Pricing Strategy	Multiplier Volatility (Mean ± SD)	Mean Absolute Multiplier Change (Mean ± SD)	Action Switch Rate (Mean ± SD)
Static Pricing	0.00000 ± 0.00000	0.00000 ± 0.00000	0.00000 ± 0.00000
Rule-Based Surge Pricing	0.02256 ± 0.00148	0.01564 ± 0.00182	0.26622 ± 0.03121
PPO-Based Pricing	0.01696 ± 0.00089	0.01185 ± 0.00126	0.23701 ± 0.02530

Source: (Research results, 2026)

Static pricing has zero values by construction because its multiplier remains fixed at 1.00 throughout each episode. Table 1 shows that PPO produced smoother pricing-policy adjustments than rule-based surge pricing. PPO achieved lower multiplier volatility (0.01696 ± 0.00089) than the surge policy (0.02256 ± 0.00148), representing a 24.83% reduction. PPO also reduced the mean absolute multiplier change by 24.21% and the action switch rate by 10.97%.

The Friedman test indicated significant global differences among the three policies for all reported stability metrics ($p < 0.001$). The Holm-adjusted paired Wilcoxon tests further confirmed that PPO had lower multiplier volatility and mean absolute multiplier change than rule-based surge pricing (both $p < 0.001$), as well as a lower action switch rate ($p = 0.00080$). These results indicate that PPO generated significantly smoother policy-level price adjustments than the rule-based surge strategy.

Table 2 summarizes the economic performance of the evaluated pricing strategies across 30 paired test episodes. Static pricing produced the lowest cumulative reward and expected revenue because it maintained the historical base price without adapting to changing demand-supply conditions. In contrast, both PPO-based pricing and rule-based surge pricing achieved substantially higher economic outcomes.

PPO-based pricing obtained the highest cumulative reward, reaching 103,316.25±3,243.99, which was slightly higher than the reward achieved by rule-based surge pricing. This result indicates that PPO was able to preserve strong economic performance while accounting for the multiplier-based stability penalty. Rule-based surge pricing produced the highest cumulative expected revenue, with 103,488.62±3,235.22, where as PPO achieved103,449.18±3,241.27. The difference was only 39.44, equivalent to approximately 0.038% of the surge-policy expected revenue.

The distinction between cumulative reward and cumulative expected revenue is important. Cumulative reward incorporates the multiplier-based stability penalty, whereas cumulative expected revenue reflects the revenue outcome before this penalty is applied. Therefore, the slightly higher cumulative reward of PPO suggests that its smoother pricing adjustments reduced the stability cost without materially sacrificing revenue. The statistical significance of these differences is reported in Table 3.

Table 2. Economic Performance Across Pricing Strategies

Pricing Strategy	Cumulative Reward (Mean ± SD)	Cumulative Expected Revenue (Mean ± SD)
Static Pricing	97,376.96 ± 2,973.89	97,376.96 ± 2,973.89
Rule-Based Surge Pricing	103,314.26 ± 3,242.19	103,488.62 ± 3,235.22
PPO-Based Pricing	103,316.25 ± 3,243.99	103,449.18 ± 3,241.27

Source: (Research results, 2026)

Note Table 2: Results are reported across 30 paired test episodes, with 300 decision steps per episode. Cumulative reward includes the multiplier-based stability penalty, whereas cumulative expected revenue is calculated before the penalty is applied. Statistical comparisons are reported in Table 3 using the Friedman test and Holm-adjusted paired Wilcoxon signed-rank tests.

Table 3 presents the inferential statistical results for the economic and policy-stability outcomes reported in Tables 1 and 2. The Friedman test identified significant global differences among static pricing, rule-based surge pricing, and PPO for all evaluated outcomes. The Holm-adjusted paired Wilcoxon tests showed that PPO significantly outperformed static pricing in cumulative reward and cumulative expected revenue. Compared with rule-based surge pricing, PPO achieved statistically

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8286

indistinguishable cumulative reward, although its expected revenue was marginally lower by 39.44, representing only 0.038% of the surge-policy expected revenue.

The stability results provide a clearer distinction between PPO and rule-based surge pricing. PPO significantly reduced multiplier volatility, mean absolute multiplier change, and action switch rate. These findings show that PPO preserved near-surge economic performance while producing smoother policy-level pricing adjustments under the same paired test scenarios.

Table 3. Inferential Statistical Results for Economic Performance and Policy Stability

Outcome	Friedman Test	Holm-Adjusted Paired Comparison	Interpretation
Cumulative Reward	($p < 0.001$)	PPO vs. Static: ($p < 0.001$); PPO vs. Surge: ($p = 0.808$)	PPO significantly outperformed static pricing, while its cumulative reward was not significantly different from rule-based surge pricing.
Cumulative Expected Revenue	($p < 0.001$)	PPO vs. Static: ($p < 0.001$); PPO vs. Surge: ($p = 0.00067$)	PPO generated significantly higher expected revenue than static pricing. Surge pricing produced marginally higher expected revenue than PPO by 39.44, equivalent to 0.038% of surge revenue.
Multiplier Volatility	($p < 0.001$)	PPO vs. Surge: ($p < 0.001$)	PPO reduced multiplier volatility by 24.83% relative to rule-based surge pricing.
Mean Absolute Multiplier Change	($p < 0.001$)	PPO vs. Surge: ($p < 0.001$)	PPO reduced the average magnitude of multiplier adjustment by 24.21%.
Action Switch Rate	($p < 0.001$)	PPO vs. Surge: ($p = 0.00080$)	PPO reduced the frequency of pricing-action changes by 10.97%.

Source: (Research results, 2026)

The inferential results indicate that PPO provides a measurable revenue-stability trade-off within the simulated pricing environment. Compared with static pricing, PPO achieved significantly higher cumulative reward and cumulative expected revenue. Compared with rule-based surge pricing, PPO produced statistically indistinguishable cumulative reward, although its cumulative expected revenue was marginally lower by 39.44, equivalent to only 0.038% of the surge-policy value.

At the same time, PPO generated significantly smoother policy-level pricing adjustments than rule-based surge pricing. As reported in Table 3, PPO reduced multiplier volatility, mean absolute multiplier change, and action switch rate under the same paired test scenarios. These findings suggest that PPO can preserve near-surge economic performance while reducing abrupt multiplier adjustments. However, the results should be interpreted within the scope of the simulated environment and do not directly establish customer trust, fairness perception, retention, or long-term platform outcomes.

CONCLUSION

This study developed and evaluated a Proximal Policy Optimization (PPO)-based dynamic pricing policy for ride-sharing under stochastic demand-supply conditions. The results show that PPO significantly outperformed static pricing in cumulative reward and cumulative expected revenue across the paired test episodes. Compared with rule-based surge pricing, PPO achieved statistically indistinguishable cumulative reward (($p = 0.808$)) while producing smoother policy-level pricing adjustments. PPO reduced multiplier volatility by 24.83%, mean absolute multiplier change by 24.21%, and action switch rate by 10.97% relative to the surge baseline. Although rule-based surge pricing generated marginally higher cumulative expected revenue, the difference was only 39.44, equivalent to 0.038% of the surge-policy value. These findings indicate that PPO can preserve near-surge economic performance while reducing abrupt multiplier adjustments within the simulated pricing environment. The findings should not be interpreted as evidence of customer trust, fairness perception, retention, or long-term platform outcomes, because these aspects were not measured. Future studies should evaluate the framework using larger time-stamped operational datasets and customer-response data from real ride-sharing settings.

REFERENCE

[1] K. Haneefa and H. Singh, "Does AI matter in marketing? Unveiling the role of AI-driven marketing in quick grocery," *Int. Rev. Retail Distrib. Consum. Res.*, pp. 1–23, 2025, doi: https://doi.org/10.1080/09593969.2025.2526482.

[2] J. Nyangon, "Smart Grid Strategies for Tackling the Duck Curve: A Qualitative Assessment of Digitalization, Battery Energy

Storage, and Managed Rebound Effects Benefits," *Energies*, vol. 18, no. 15, p. 3988, 2025, doi: https://doi.org/10.3390/en18153988.

[3] N. Alamsyah, A. P. Kurniati, and others, "Airfare Fluctuation Analysis with Event and Sentiment Features by Stacking Ensemble Model," in *2024 Ninth International Conference on Informatics and Computing (ICIC)*, IEEE, 2024, pp. 1–6. doi: 10.1109/ICIC64337.2024.10957538.

[4] L. Ming, T. Tunca, Y. Xu, and W. Zhu, "Market Formation, Pricing, and Value Generation in Ride-Hailing Services," *Manuf. Serv. Oper. Manag.*, vol. 27, no. 5, pp. 1551–1570, 2025, doi: https://doi.org/10.1287/msom.2022.0502.

[5] J. Li and B. Chen, "A Deep Q-Learning Optimization Framework for Dynamic Pricing in E-Commerce," in *Proceedings of the 2025 4th International Conference on Cyber Security, Artificial Intelligence and the Digital Economy*, 2025, pp. 367–371. doi: https://doi.org/10.1145/3729706.37297.

[6] J. Y. Yang, "The Future of AI-Enabled Pricing," in *Reimagine Pricing: How AI is Changing Everything*, Springer, 2025, pp. 131–142. doi: https://doi.org/10.1007/978-3-031-90418-9_5.

[7] N. Alamsyah, A. P. Kurniati, and others, "Event Detection Optimization Through Stacking Ensemble and BERT Fine-Tuning for Dynamic Pricing of Airline Tickets," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3466270.

[8] Y. Kayikci, S. Demir, S. K. Mangla, N. Subramanian, and B. Koc, "Data-driven optimal dynamic pricing strategy for reducing perishable food waste at retailers," *J. Clean. Prod.*, vol. 344, p. 131068, 2022, doi: https://doi.org/10.1016/j.jclepro.2022.131068.

[9] T. A. Syed, H. Aslam, Z. A. Bhatti, F. Mehmood, and A. Pahuja, "Dynamic pricing for perishable goods: A data-driven digital transformation approach," *Int. J. Prod. Econ.*, vol. 277, p. 109405, 2024, doi: https://doi.org/10.1016/j.ijpe.2024.109405.

[10] R. Y. Chenavaz and S. Dimitrov, "Artificial intelligence and dynamic pricing: a systematic literature review," *J. Appl. Econ.*, vol. 28, no. 1, p. 2466140, 2025, doi: https://doi.org/10.1080/15140326.2025.2466140.

[11] X. Li, C. Schmidt, D. Gammelli, and F. Rodrigues, "Learning Joint Rebalancing and Dynamic Pricing Policies for Autonomous Mobility-on-Demand," *IEEE Trans. Intell. Transp. Syst.*, 2025, doi: 10.1109/TITS.2025.3582639.

[12] D. Wang, X. Chen, X. Liu, Y. Li, Z. Piao, and H. Li, "Dynamic Tariff Adjustment for Electric Vehicle Charging in Renewable-Rich Smart Grids: A Multi-Factor Optimization Approach to Load Balancing and Cost Efficiency," *Energies*, vol. 18, no. 16, p. 4283, 2025, doi: https://doi.org/10.3390/en18164283.

[13] U. Y. Hasanah and R. Rino, "Pricing and Adaptation Strategies in Market Dynamics: A Systematic Literature Review," *Electron. J. Educ. Soc. Econ. Technol.*, vol. 6, no. 1, 2025, doi: https://doi.org/10.33122/ejeset.v6i1.497.

[14] L. Cheng, M. Li, C. Tan, P. Huang, M. Zhang, and R. Sun, "Computational game-theoretic models for adaptive urban energy systems: A comprehensive review of algorithms, strategies, and engineering applications," *Arch. Comput. Methods Eng.*, pp. 1–78, 2025, doi: https://doi.org/10.1007/s11831-025-10364-y.

[15] X. Guo and L. Zhang, "Dynamic Pricing Models in E-Commerce: Exploring Machine Learning Techniques to Balance Profitability and Customer Satisfaction," *IEEE Access*, 2025, doi: 10.1109/ACCESS.2025.3563371.

[16] B. Jin and X. Xu, "Machine learning-based forecasts of residential property prices in Hangzhou city, Zhejiang province, China," *Neural Comput. Appl.*, vol. 37, no. 6, pp. 4971–4988, 2025, doi: https://doi.org/10.1007/s00521-024-10726-w.

[17] L. Zheng and L. Tan, "A decentralized scheme for multi-user edge computing task offloading based on dynamic pricing," *Peer--Peer Netw. Appl.*, vol. 18, no. 2, p. 91, 2025, doi: https://doi.org/10.1007/s12083-025-01904-1.

[18] A. Pagliaro, "Artificial intelligence vs. efficient markets: A critical reassessment of predictive models in the big data era," *Electronics*, vol. 14, no. 9, p. 1721, 2025, doi: https://doi.org/10.3390/electronics14091721.

[19] M. Bichler, J. Durmann, and M. Oberlechner, "Algorithmic Pricing and Algorithmic Collusion: M. Bichler et al.," *Bus. Inf. Syst. Eng.*, pp. 1–9, 2025, doi:

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8286

https://doi.org/10.1007/s12599-025-00965-z.

[20] P. Michailidis, I. Michailidis, and E. Kosmatopoulos, "Reinforcement Learning for Electric Vehicle Charging Management: Theory and Applications," *Energies*, vol. 18, no. 19, p. 5225, 2025, doi: https://doi.org/10.3390/en18195225.

[21] S. Giannelos, "Reinforcement Learning in Energy Finance: A Comprehensive Review.," *Energ. 19961073*, vol. 18, no. 11, 2025, doi: 10.3390/en18112712.

[22] P. Michailidis, I. Michailidis, C. R. Lazaridis, and E. Kosmatopoulos, "Traffic Signal Control via Reinforcement Learning: A Review on Applications and Innovations," *Infrastructures*, vol. 10, no. 5, p. 114, 2025, doi: https://doi.org/10.3390/infrastructures10050114.

[23] Z. Zhou *et al.*, "A Transformer-Based Reinforcement Learning Framework for Sequential Strategy Optimization in Sparse Data," *Appl. Sci.*, vol. 15, no. 11, p. 6215, 2025, doi: https://doi.org/10.3390/app15116215.

[24] A. Kavoosi, R. Tavakkoli-Moghaddam, H. Sajedi, N. Tajik, and K. Tafakkori, "Dynamic pricing and inventory control of perishable products by a deep reinforcement learning algorithm," *Expert Syst. Appl.*, vol. 291, p. 128570, 2025, doi: https://doi.org/10.1016/j.eswa.2025.128570.

[25] I. Poulaki, N. I. Koufodontis, and S. Papadimitriou, "Airline revenue management, distribution and passengers: market trends in a technology driven triangle," *Worldw. Hosp. Tour. Themes*, vol. 17, no. 1, pp. 35–47, 2025, doi: https://doi.org/10.1108/WHATT-12-2024-0304.

[26] H. Xu, A. Zhang, Q. Wang, Y. Hu, F. Fang, and L. Cheng, "Quantum Reinforcement Learning for real-time optimization in Electric Vehicle charging systems," *Appl. Energy*, vol. 383, p. 125279, 2025, doi: https://doi.org/10.1016/j.apenergy.2025.125279.