Laporan Analisis Reinforcement Learning: Q-Learning vs. SARSA

1. Cara Kerja Q-Learning dan SARSA

Q-Learning adalah algoritma *Reinforcement Learning off-policy* yang mempelajari kebijakan optimal dengan memilih aksi terbaik di masa depan, tanpa mempedulikan aksi yang diambil saat eksplorasi. Dengan memperbarui *Q-table* menggunakan *Q-value* maksimum dari semua aksi di keadaan berikutnya, membuatnya cenderung mencari jalur terpendek, meski kadang berisiko.

SARSA (*State-Action-Reward-State-Action*) adalah algoritma *on-policy* yang memperbarui isi *Q-table* berdasarkan kebijakan yang dijalankan, termasuk aksi random dari eksplorasi (misalnya, *epsilon-greedy*). SARSA menggunakan *Q-value* dari aksi yang diambil di *state* berikutnya, sehingga lebih hati-hati karena mempertimbangkan risiko aksi yang random.

2. Perbandingan Hasil di Wumpus World

Kebijakan (Policy) Final dan Jalur yang Ditempuh:
Setelah pelatihan di lingkungan *Wumpus World*, kedua agen berhasil menemukan kebijakan untuk mengambil emas dan kembali ke titik awal. Namun, jalur optimal yang ditempuh sedikit berbeda, mencerminkan perbedaan strategi belajar kedua algoritma.

- Jalur Optimal Q-Learning: (3, 0) → (2, 0) → (2, 1) → (2, 2) [Emas] → (2, 1) → (3, 1) → (3, 0) [Pulang]
- Jalur Optimal SARSA: (3, 0) → (3, 1) → (2, 1) → (2, 2) [Emas] → (2, 1) → (2, 0) → (3, 0) [Pulang]

Analisis Perbandingan:
Kedua agen berhasil menyelesaikan tugas, membuktikan bahwa Q-Learning dan SARSA mampu memecahkan masalah *Wumpus World*. Perbedaan utama ada pada rute pulang setelah mengambil emas:

- Q-Learning (*Off-policy*/Agresif): Agen Q-Learning memilih rute pulang (2, 1) → (3, 1) → (3, 0), yang merupakan salah satu jalur terpendek dan aman. Karena *off-policy*, Q-Learning fokus pada aksi dengan nilai *Q-value* tertinggi, mengabaikan risiko eksplorasi, sehingga cenderung agresif mencari jalur efisien.
- SARSA (*On-policy*/Konservatif): Agen SARSA memilih rute pulang (2, 1) → (2, 0) → (3, 0). Meskipun juga efisien, rute ini menunjukkan kecenderungan SARSA untuk lebih hati-hati, karena mempertimbangkan aksi acak selama eksplorasi (*epsilon-greedy*). Rute ini mungkin dianggap lebih aman berdasarkan pengalaman belajarnya, terutama di lingkungan dengan risiko seperti *pit*.

Perbedaan jalur, meski kecil, menyoroti bahwa SARSA lebih konservatif dibandingkan Q-Learning, yang lebih agresif mencari jalur terpendek.