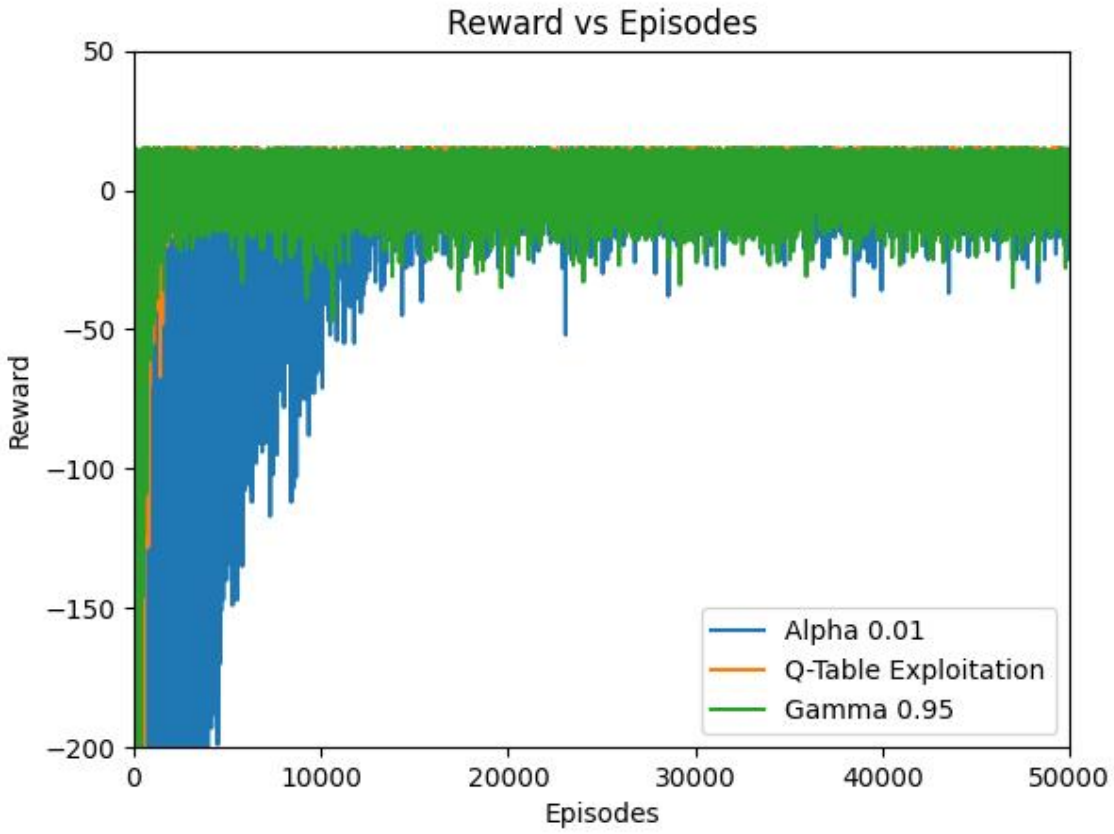
Reinforcement Learning - Hiperparâmetros

Autor: Renato Laffranchi Falcão

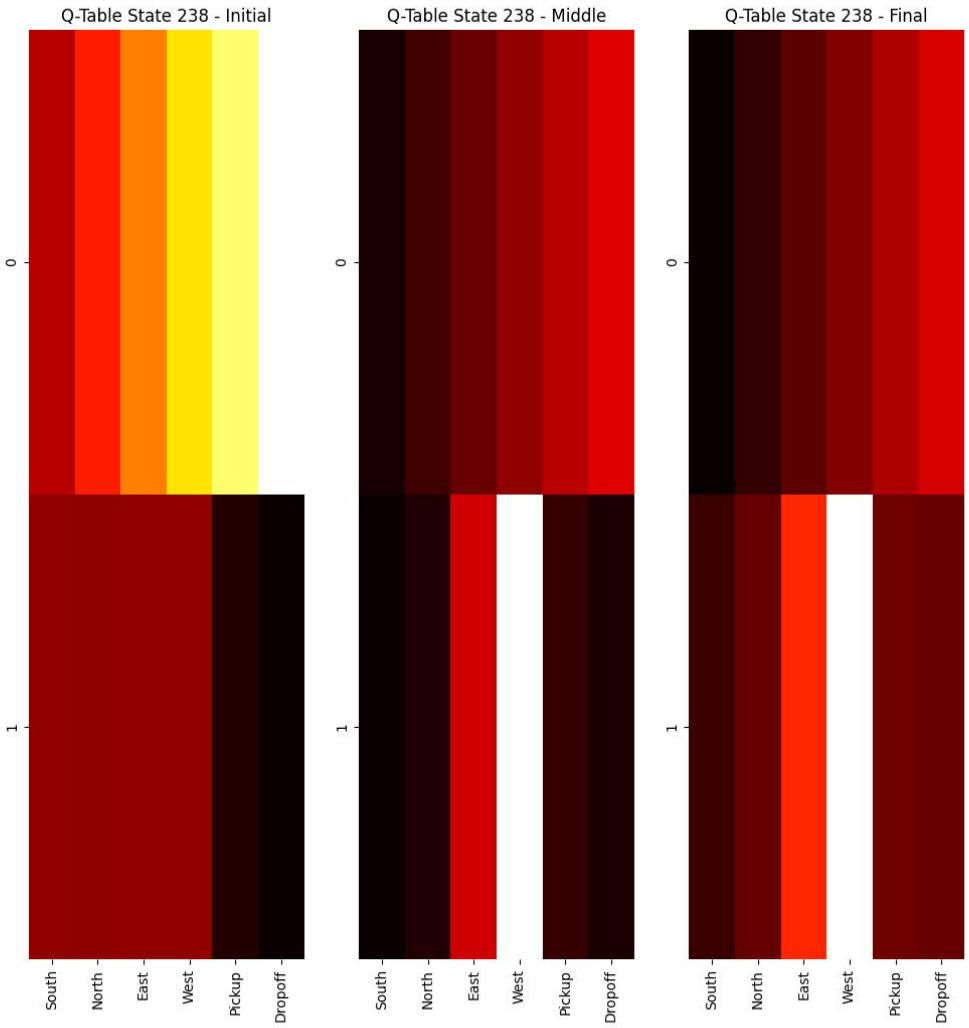


1. **Gráfico de Recompensa vs Episódios**:
   Este gráfico mostra o desempenho de um modelo de aprendizado por reforço ao longo de episódios de treinamento. Cada ponto no gráfico representa a recompensa total obtida por episódio. O eixo X representa os episódios, e o eixo Y representa a recompensa. Existem três linhas que representam diferentes aspectos do modelo de aprendizado:

   - **Alpha 0.01**: Esta linha pode representar a taxa de aprendizado do modelo. Um valor de alpha de 0.01 indica que o modelo está aprendendo lentamente, incorporando apenas 1% da nova informação a cada atualização da Q-table.

- **Q-Table Exploitation**: Essa linha verde indica a recompensa obtida ao explorar o conhecimento atual da Q-table, sem explorar novas ações. Isso mostra como o modelo se comporta ao usar as melhores ações conhecidas.
- **Gamma 0.95**: Representa o fator de desconto, que determina a importância das recompensas futuras. Um valor de 0.95 sugere que o modelo valoriza bastante as recompensas futuras, mas não tanto quanto as imediatas.

A imagem parece mostrar uma estabilização das recompensas ao longo do tempo, indicando que todos modelo está tendo muita dificuldade em convergir, seja qual for a variação de hiperparâmetro analisada, porque é perceptível que as recompensas variam muito, indicando que o modelo não consegue aprender com eficiência para nenhuma das variações de hiperparâmetro.



1. **Heatmap da Q-Table**:
   A segunda imagem é um conjunto de três heatmaps que representam o estado da Q-Table em três momentos diferentes do treinamento: inicial, meio e final. O heatmap mostra os valores para um estado específico (Estado 238) e para diferentes ações possíveis. A cor de cada célula no heatmap representa o valor Q para o par estado-ação correspondente, com cores mais claras indicando valores mais altos e cores mais escuras indicando valores mais baixos.

   - **Q-Table State 238 - Initial**: Mostra os valores iniciais, que podem ser aleatórios ou baseados em alguma inicialização prévia.

- **Q-Table State 238 - Middle**: Mostra os valores da Q-table na metade do treinamento, onde já podemos ver alguma aprendizagem acontecendo, com valores sendo ajustados com base nas recompensas recebidas.
- **Q-Table State 238 - Final**: Mostra os valores após o treinamento ser concluído. Idealmente, esses valores representam uma política bem definida onde o modelo aprendeu quais ações são melhores em dado estado.