Reinforcement Learning CIA-2

Create a 100x100 grid with obstacles in between 2 random points. Build an MDP based RL agent to optimise both policies and actions at every state. Benchmark DP method with other RL solutions for the same problem.

Aim:-

Markov Decision Process (MDP)-based Reinforcement Learning (RL) agent for a 100x100 grid with obstacles. The goal is to train an agent to navigate from a random start point to a goal point, optimizing its policy and actions while benchmarking different solution methods.

Procedure:-

1.Grid Setup:

- Create a 100x100 grid, initializing it with free cells and adding obstacles randomly.
- Define a start and end point at random locations on the grid, ensuring that there's a valid path from the start to the goal.

2. **MDP Definition**:

- **States**: Each cell in the grid is a state S={(x,y)|x,y∈[0,99]}S = \{(x, y) | x, y \in [0, 99]\}S={(x,y)|x,y∈[0,99]}.
- **Actions**: The agent can move up, down, left, or right if there are no obstacles.
- **Rewards**: Assign a reward of +1 for reaching the goal, 0 for other cells, and -1 for attempting to move into an obstacle or going out of bounds.
- **Transitions**: Define the state transitions based on chosen actions and the presence of obstacles.

3. **Policy Optimization Methods**:

- **Q-learning**: A model-free RL method to learn the optimal action-value function.
- **SARSA**: An on-policy algorithm, can be compared to Q-learning to see how following the policy impacts learning.

4. **Algorithm Benchmarking**:

- Run each algorithm (Q-learning, SARSA) on the same grid environment, tracking metrics such as:
    - Convergence time.
    - Success rate (agent reaching the goal).
    - Average reward collected over episodes.

5. **Implementation Steps**:

    - For Q-learning and SARSA: Initialize Q-tables or function approximators and update them based on sampled actions.

6. **Performance Evaluation**:

    - Record the time and episode count required for each method to reach a defined success criterion.
    - Track the agent's performance over multiple episodes to evaluate consistency and reliability.