Problem 2

1. Reinforcement Learning VS Supervised Learning

First Reinforcement learning is about sequential decision making. What that means is, given the current input, you make a decision, and the *next input depends on your decision*. In supervised learning, the decisions you make, either in a batch setting, or in an online setting, do not affect what you see in the future. This is the fundamental difference between supervised learning and reinforcement learning.

Second supervised learning is when a model learns from a labeled dataset with guidance Whereas reinforcement learning is when a machine or an agent interacts with its environment, performs actions, and learns by a trial-and-error method.

1. Value Iteration Vs Policy Iteration

Policy iteration and value iteration are both dynamic programming algorithms that find an optimal policy $\boldsymbol{\pi}$ in a reinforcement learning environment. They both employ variations of Bellman updates and exploit one-step look-ahead:

Policy Iteration	Value Iteration
Starts with a random policy	Starts with a random value function
Algorithm is more complex	Algorithm is simpler
Guaranteed to converge	Guaranteed to converge
Cheaper to compute	More expensive to compute
Requires few iterations to converge	Requires more iterations to converge
Faster	Slower

In policy iteration, we start with a fixed policy. Conversely, in value iteration, we begin by selecting the value function. Then, in both algorithms, we iteratively improve until we reach convergence.

The policy iteration algorithm updates the policy. The value iteration algorithm iterates over the value function instead. Still, both algorithms implicitly update the policy and state value function in each iteration.

In each iteration, the policy iteration function goes through two phases. One phase evaluates the policy, and the other one improves it. The value iteration function covers these two phases by taking a maximum over the utility function for all possible actions.

The value iteration algorithm is straightforward. It combines two phases of the policy iteration into a single update operation. However, the value iteration function runs through all possible actions at once to find the maximum action value. Subsequently, the value iteration algorithm is computationally heavier.

Both algorithms are guaranteed to converge to an optimal policy in the end. Yet, the policy iteration algorithm converges within fewer iterations. As a result, the policy iteration is reported to conclude faster than the value iteration algorithm.