Câu 1:

Câu hỏi: Hãy giải thích sự khác biệt giữa Học có giám sát (Supervised Learning) và Học không giám sát (Unsupervised Learning).

Đáp án:

- **Học có giám sát (Supervised Learning)** là phương pháp học máy trong đó mô hình được huấn luyện với tập dữ liệu có nhãn. Điều này có nghĩa là mỗi đầu vào (input) sẽ có một đầu ra mong muốn (output) đã biết trước. Ví dụ: phân loại email thành spam hoặc không spam.
- **Học không giám sát (Unsupervised Learning)** là phương pháp học máy mà mô hình không có nhãn đầu ra rõ ràng. Mục tiêu là tìm ra các mẫu hoặc cấu trúc ẩn trong dữ liệu, chẳng hạn như phân cụm khách hàng theo hành vi mua sắm.

Câu 2:

Câu hỏi: Hãy mô tả và giải thích vai trò của `Confusion Matrix` trong đánh giá mô hình phân loại.

Đáp án:

- `Confusion Matrix` là một bảng gồm 4 ô dùng để đánh giá hiệu suất của mô hình phân loại:
  - **True Positive (TP):** Dự đoán đúng mẫu thuộc lớp dương.
  - **False Positive (FP):** Dự đoán sai mẫu không thuộc lớp dương thành lớp dương.
  - **True Negative (TN):** Dự đoán đúng mẫu thuộc lớp âm.
  - **False Negative (FN):** Dự đoán sai mẫu thuộc lớp dương thành lớp âm.
- `Confusion Matrix` giúp tính toán các chỉ số quan trọng như:
  - **Accuracy (Độ chính xác)** = (TP + TN) / (TP + TN + FP + FN)
  - **Precision (Độ chính xác của lớp dương)** = TP / (TP + FP)
  - **Recall (Độ bao phủ của lớp dương)** = TP / (TP + FN)
  - **F1-score** = 2 * (Precision * Recall) / (Precision + Recall)

Câu 3:

Câu hỏi: Hãy giải thích khái niệm `Overfitting` và cách giảm thiểu nó trong mô hình học máy.

Đáp án:

- `Overfitting` là hiện tượng mô hình học máy học quá kỹ trên tập huấn luyện đến mức nó không thể tổng quát hóa tốt trên dữ liệu mới. Mô hình ghi nhớ dữ liệu thay vì học quy luật, dẫn đến hiệu suất kém trên tập kiểm tra.

- Các cách giảm thiểu `Overfitting`:
    1. **Thu thập thêm dữ liệu:** Mô hình có nhiều mẫu hơn để học các quy luật tổng quát.
    2. **Regularization (Chuẩn hóa):** Sử dụng L1/L2 Regularization (Lasso, Ridge) để giảm trọng số của các đặc trưng không quan trọng.
    3. **Pruning (Cắt tỉa):** Nếu sử dụng cây quyết định, có thể áp dụng kỹ thuật cắt tỉa để giảm độ phức tạp của cây.
    4. **Dropout (Đối với mạng neuron):** Tắt ngẫu nhiên một số neuron trong quá trình huấn luyện để tránh quá khớp.
    5. **Cross-validation (Kiểm định chéo):** Dùng kỹ thuật k-fold cross-validation để đánh giá mô hình trên nhiều tập con khác nhau.