BÁO CÁO TUẦN 1

1. Phân nhóm các thuật toán Machine Learning

Có hai cách phổ biến để phân loại các thuật toán Machine Learning:

Dựa trên phương thức học (Learning style):
- Học có giám sát (Supervised Learning)
- Học không giám sát (Unsupervised Learning)
- Học tăng cường (Reinforcement Learning)

Dựa trên chức năng (Function):
- Classification
- Regression
- Clustering
- Dimensionality Reduction
- Recommendation Systems

2. Linear Regression

Linear Regression (Hồi quy tuyến tính) là thuật toán dùng để dự đoán giá trị liên tục dựa trên mối quan hệ tuyến tính giữa các biến đầu vào (features) và biến đầu ra (target). Mục tiêu là tìm một hàm tuyến tính sao cho sai số giữa giá trị dự đoán và giá trị thực tế là nhỏ nhất.

Công thức mô hình:

$y = w1*x1 + w2*x2 + ... + wn*xn + b$

Trong đó:

- `w1, w2, ...,` wn là các hệ số (weights) đại diện cho tầm quan trọng của mỗi feature.
- b là hệ số chặn (bias/intercept).
- `x1, x2, ...,` xn là các giá trị của features.

Hàm mất mát (Loss function):

- Mean Squared Error (MSE) thường được sử dụng:
- Mục tiêu: tối thiểu hóa MSE bằng cách cập nhật `weights` và `bias` thông qua Gradient Descent.

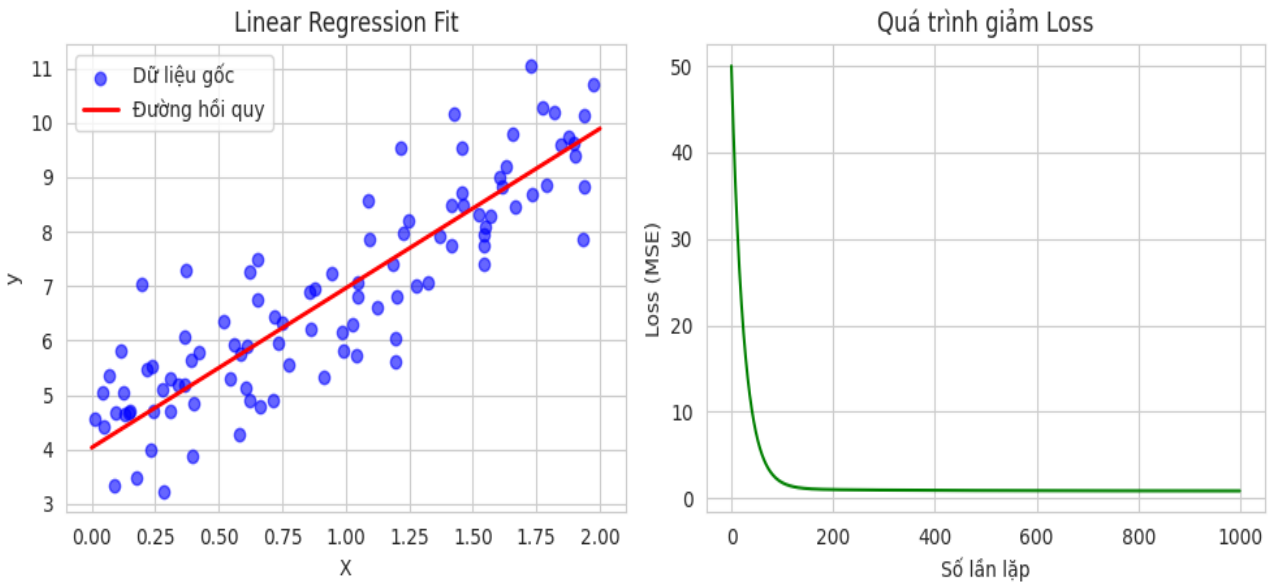
Các bước Gradient Descent:

1. Khởi tạo weights và bias (thường là zeros).
2. Tính giá trị dự đoán: `y_pred = X.dot(weights) + bias`.
3. Tính gradient của loss đối với weights và bias.
4. Cập nhật weights và bias:

weights -= learning_rate * gradient_weights

bias -= learning_rate * gradient_bias

5. Lặp lại quá trình cho đến khi đạt số iteration hoặc hàm mất mát hội tụ.



3. K-Nearest Neighbors (KNN)

KNN là thuật toán giám sát dùng cho phân loại hoặc hồi quy, dựa trên khoảng cách giữa điểm dữ liệu mới và các điểm dữ liệu đã biết.

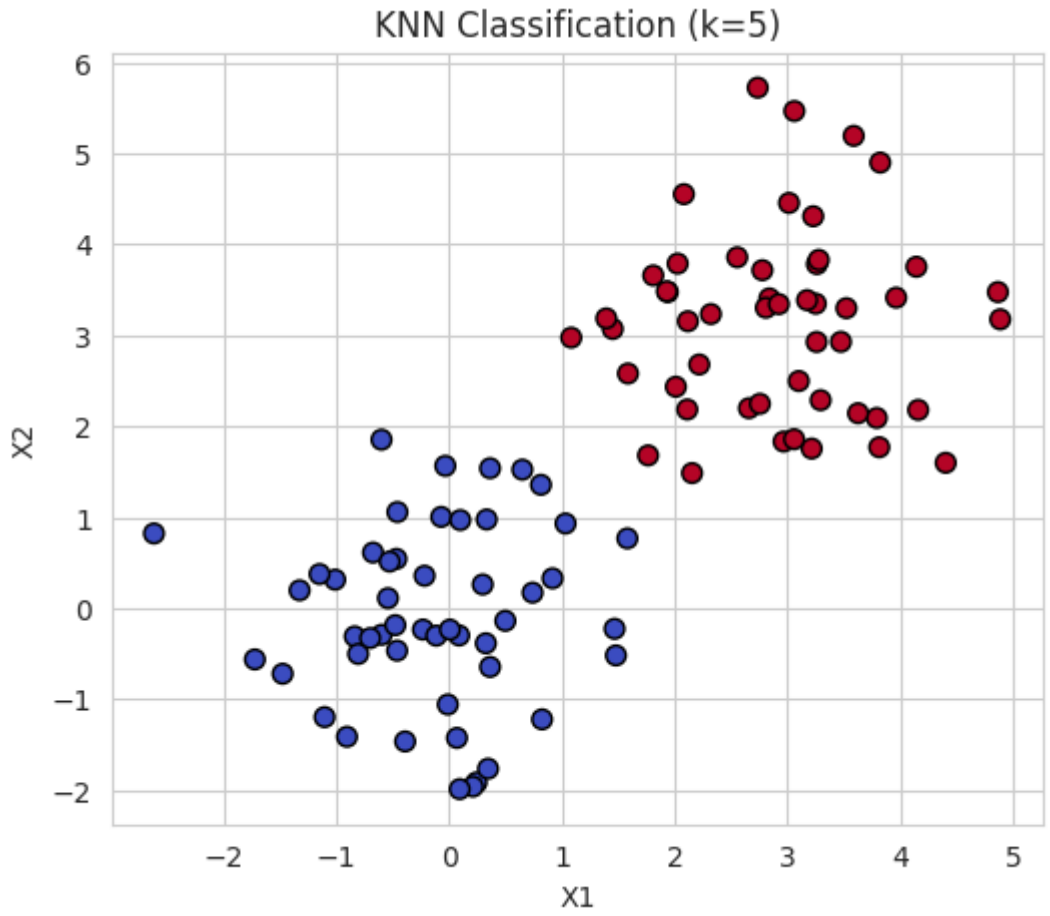
Các bước thực hiện:

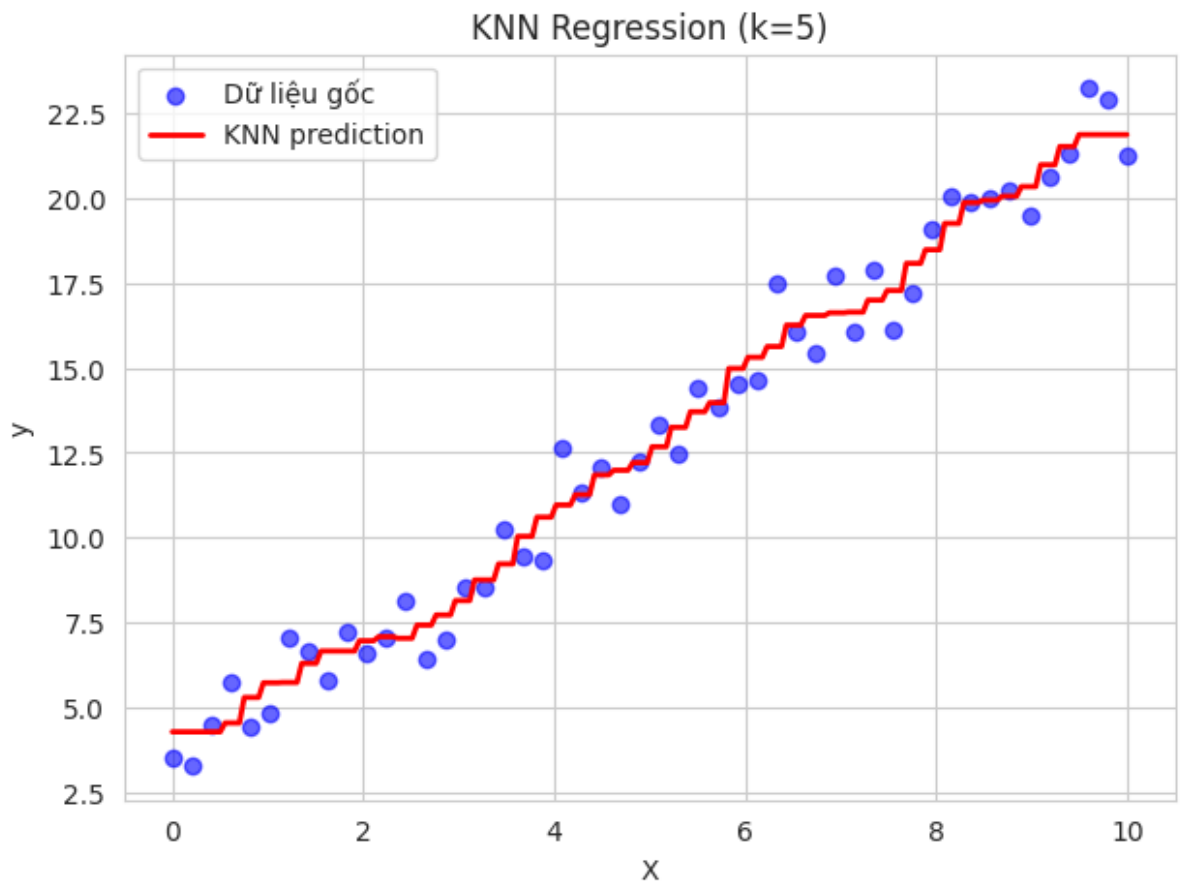
1. Chọn số k (số lượng hàng xóm gần nhất).
2. Tính khoảng cách từ điểm cần dự đoán đến tất cả điểm trong tập huấn luyện.
3. Chọn k điểm gần nhất.
4. Dự đoán:
    - Classification: nhãn chiếm đa số trong k điểm gần nhất.
    - Regression: giá trị trung bình của các nhãn trong k điểm gần nhất.

Khoảng cách thường dùng:

- Euclidean distance: `sqrt(Σ(x_i - x'_i)^2)`

- Manhattan distance: $\Sigma|x_i - x'_i|$





4. K-means Clustering

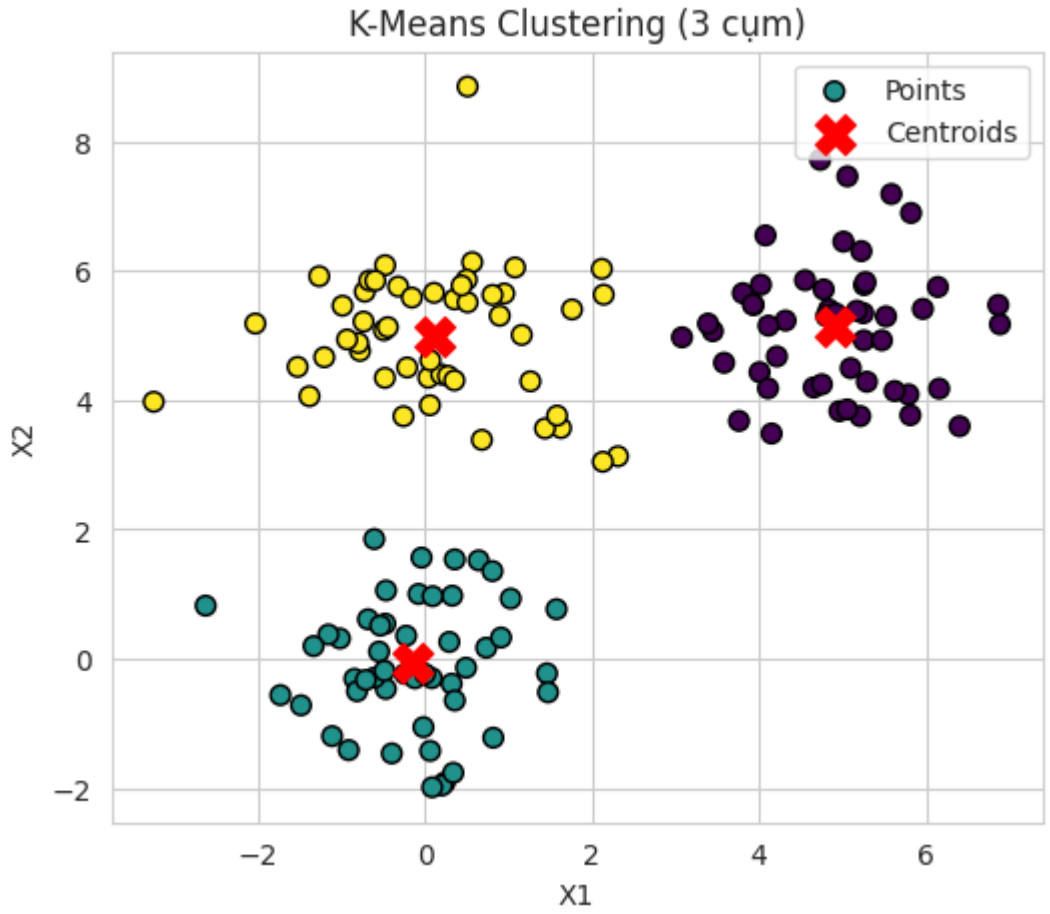
K-Means Clustering là thuật toán phân cụm dựa trên khoảng cách giữa các điểm dữ liệu, với mục tiêu nhóm các điểm dữ liệu thành K cụm sao cho tổng khoảng cách từ các điểm đến tâm cụm gần nhất là nhỏ nhất.

- **Centroid**: Tâm của mỗi cụm, là trung bình tọa độ của tất cả điểm trong cụm.

- **K**: Số lượng cụm mong muốn (do người dùng chọn trước).

Quy trình thực hiện:

1. Chọn số lượng cluster k.
2. Khởi tạo ngẫu nhiên k tâm centroids.
3. Lặp lại cho tới khi hội tụ hoặc đạt max iteration:
    - Gán mỗi điểm dữ liệu vào cluster có centroid gần nhất.
    - Cập nhật centroid mới bằng trung bình các điểm thuộc cluster.



5. Đánh giá mô hình Machine Learning

Sau khi huấn luyện mô hình, việc đánh giá hiệu suất là rất quan trọng để đảm bảo mô hình hoạt động tốt trên dữ liệu

thực tế.
Các chỉ số đánh giá phổ biến bao gồm:

1. Linear Regression:
- Mean Squared Error (MSE): Trung bình bình phương sai số giữa giá trị thực tế và giá trị dự đoán.
- R-squared (R^2): Tỷ lệ phương sai được giải thích bởi mô hình. Giá trị gần 1 cho thấy mô hình phù hợp tốt.

2. K-Nearest Neighbors (KNN):
- Classification: Accuracy (độ chính xác), Precision, Recall, F1-score.
- Regression: Mean Squared Error (MSE), Mean Absolute Error (MAE), R^2.

3. K-means Clustering:
- Sum of Squared Errors (SSE): Tổng bình phương khoảng cách từ các điểm dữ liệu tới tâm cụm, càng nhỏ càng tốt.
- Silhouette Score: Đo mức độ tách biệt và chặt chẽ của các cụm, giá trị từ -1 đến 1.

Việc lựa chọn chỉ số đánh giá phù hợp tùy thuộc vào loại bài toán.