CS313.N22 – Khai Thác Dữ Liệu và Ứng Dụng

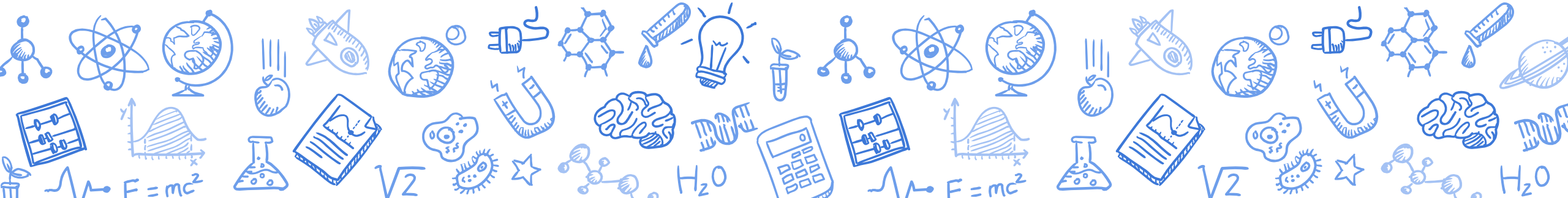
Đề tài

Dự đoán kết quả học tập của các môn học mà sinh viên lựa chọn cho học kỳ tiếp theo

GVHD: Nguyễn Thị Anh Thư

Nhóm 3

Hồ Thanh Tịnh	*20520813*
Lê Nguyễn Bảo Hân	*20520174*
Nguyễn Văn Đức Ngọc	*20521666*
Bùi Nguyễn Anh Trung	*20520332*
Nguyễn Trần Minh Anh	*20520394*



Nội dung

1. Giới thiệu đề tài

2 Nội dung thực hiện

- *Tiền xử lý dữ liệu*
- *Các phương pháp đề xuất*

3. Thực nghiệm

4. Kết luận và hướng phát triển

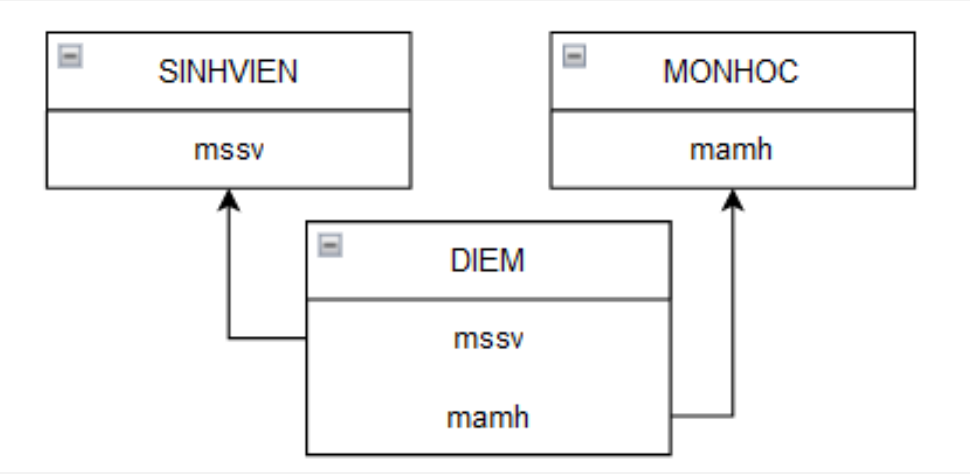
Giới thiệu đề tài

- Lộ trình học không phù hợp gây ảnh hưởng đến hiệu quả học tập và tỷ lệ tốt nghiệp đúng hạn
- Xây dựng mô hình dự đoán **điểm các môn học** mà sinh viên lựa chọn trong học kỳ tiếp theo
- Ba hướng tiếp cận: mô hình máy học, mô hình neural network và mô hình hệ thống khuyến nghị

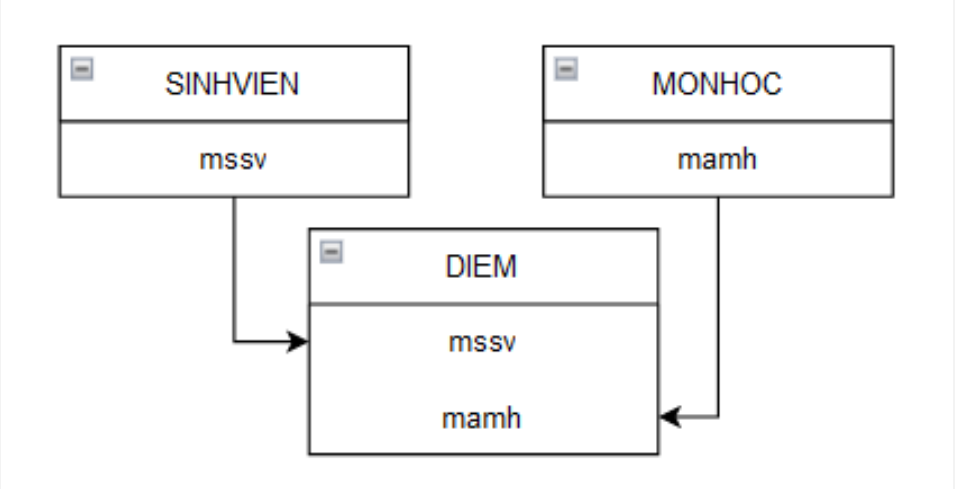
Tiền xử lý dữ liệu

Cấu trúc dữ liệu cho các phương pháp

- Tuy 3 phương pháp được sử dụng để thực nghiệm đều hoạt động tốt trên các cấu trúc khác nhau và thuộc tính khác nhau. Nhưng điểm chung giữa chúng chính là 3 nhóm thông tin cơ bản cần thiết để huấn luyện mô hình: thông tin về điểm, sinh viên và môn học.



Recommendation System

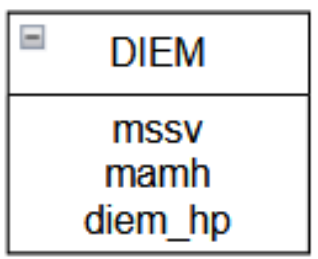


Machine Learning & Neural Network

Tiền xử lý dữ liệu

Thông tin điểm

- Bảng DIEM được dựa trên file diem_Thu và là cơ sở xây dựng nên bộ dữ liệu.
- Vì là cơ sở nên bảng này có vai trò quan trọng trong hình thành dữ liệu:
  - Với phương pháp RS, là bảng liên kết với các bảng sinh viên và môn học.
  - Với phương pháp ML và NN, là bảng tiền đề xây dựng dữ liệu.

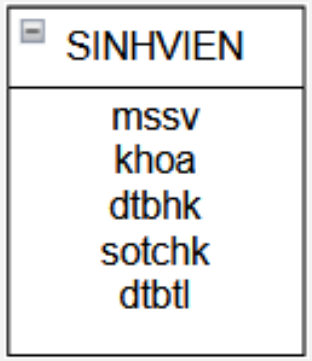


Recommendation System

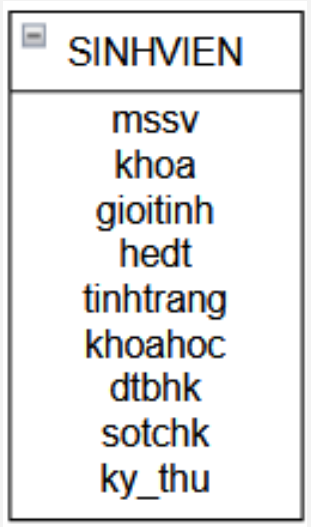
Machine Learning & Neural Network

Tiền xử lý dữ liệu

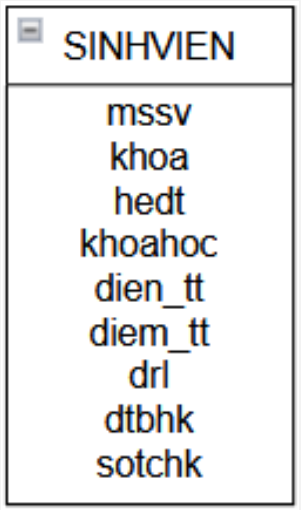
Thông tin sinh viên



Recommendation System



Machine Learning



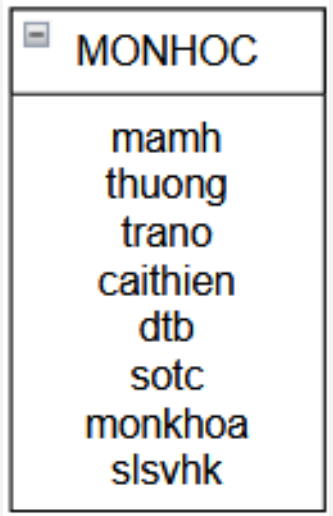
Neural Network

- Các thông tin cơ bản của sinh viên hầu hết đều được tích hợp thông qua bảng 01.sinhvien bằng 'mssv'.
- Chỉ riêng các cột sotchk và dtbhk được tổng hợp trên chính file diem_Thu và thống kế theo từng học kỳ, sau đó tích hợp vào bảng.

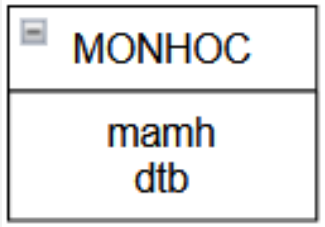
Tiền xử lý dữ liệu

Thông tin môn học

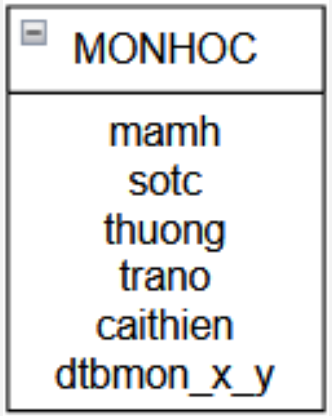
- Các cột thuong, trano, caithien lần lượt chính là sinh viên đăng ký môn học với trạng thái 1, 2, 3. Các trạng thái được thống kê từ file diem_Thu và nhóm bằng 'mamh'.
- Và điểm trung bình có được từ việc nhóm dữ liệu theo 'mamh' và hàm mean.



Recommendation System



Machine Learning



Neural Network

➔ **Nếu thông tin sinh viên dùng để phản ánh trình độ sinh viên thì thông tin môn học dùng để phản ánh độ khó của môn học.**

Tiền xử lý dữ liệu

Phân chia dữ liệu

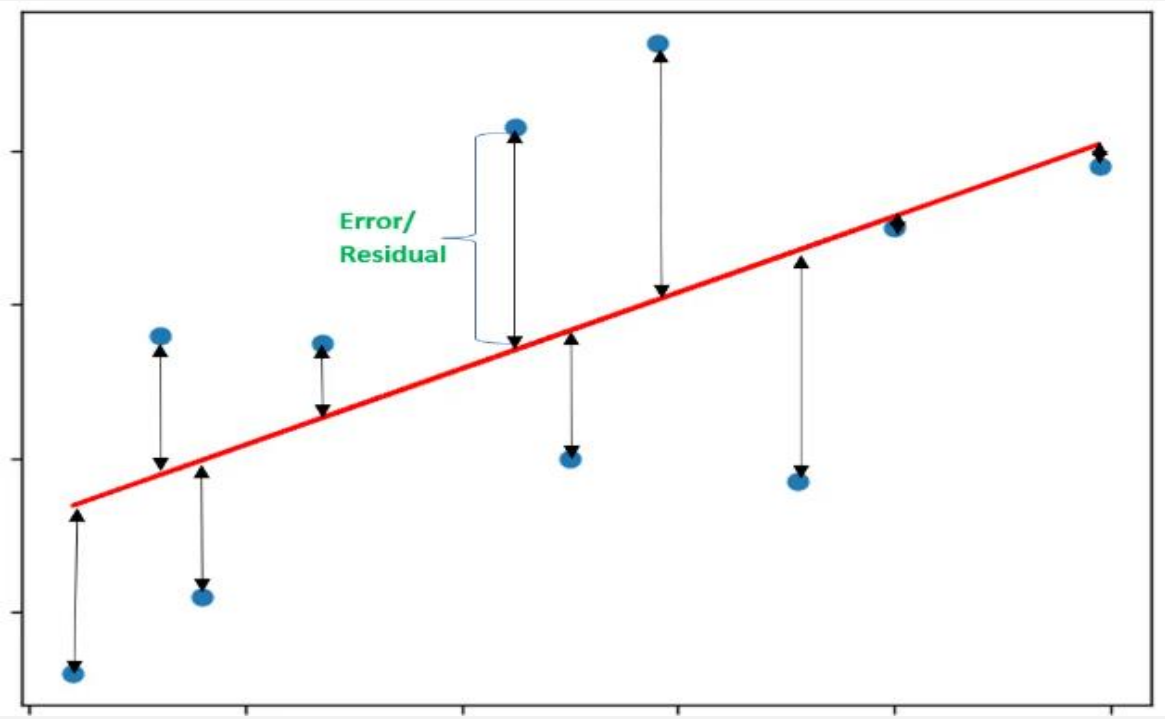
- Phân chia dữ liệu:
  - Tập train: gồm dữ liệu trước năm 2022.
  - Tập test: dữ liệu điểm sinh viên có được vào năm 2022.

```
train_data = data_filled.loc[data['namhoc'] < 2022]
test_data = data_filled.loc[data['namhoc'] == 2022]

X_train = train_data.drop(['diem_hp'], axis=1)
y_train = train_data['diem_hp']

X_test = test_data.drop(['diem_hp'], axis=1)
y_test = test_data['diem_hp']
```

- Sau khi phân chia, tiến hành tạo dữ liệu đầu vào và đầu ra. Với cột ‘diem_hp’ là dữ liệu đầu ra và các cột còn lại là thông tin đầu vào.

Mô hình hồi quy

Linear Regression và các biến thể



$$Y = w^T X$$

$$Y^{(i)} = w_0 + w_1 * x_1^{(i)} + w_2 * x_2^{(i)} + .. + w_n * x_n^{(i)}$$

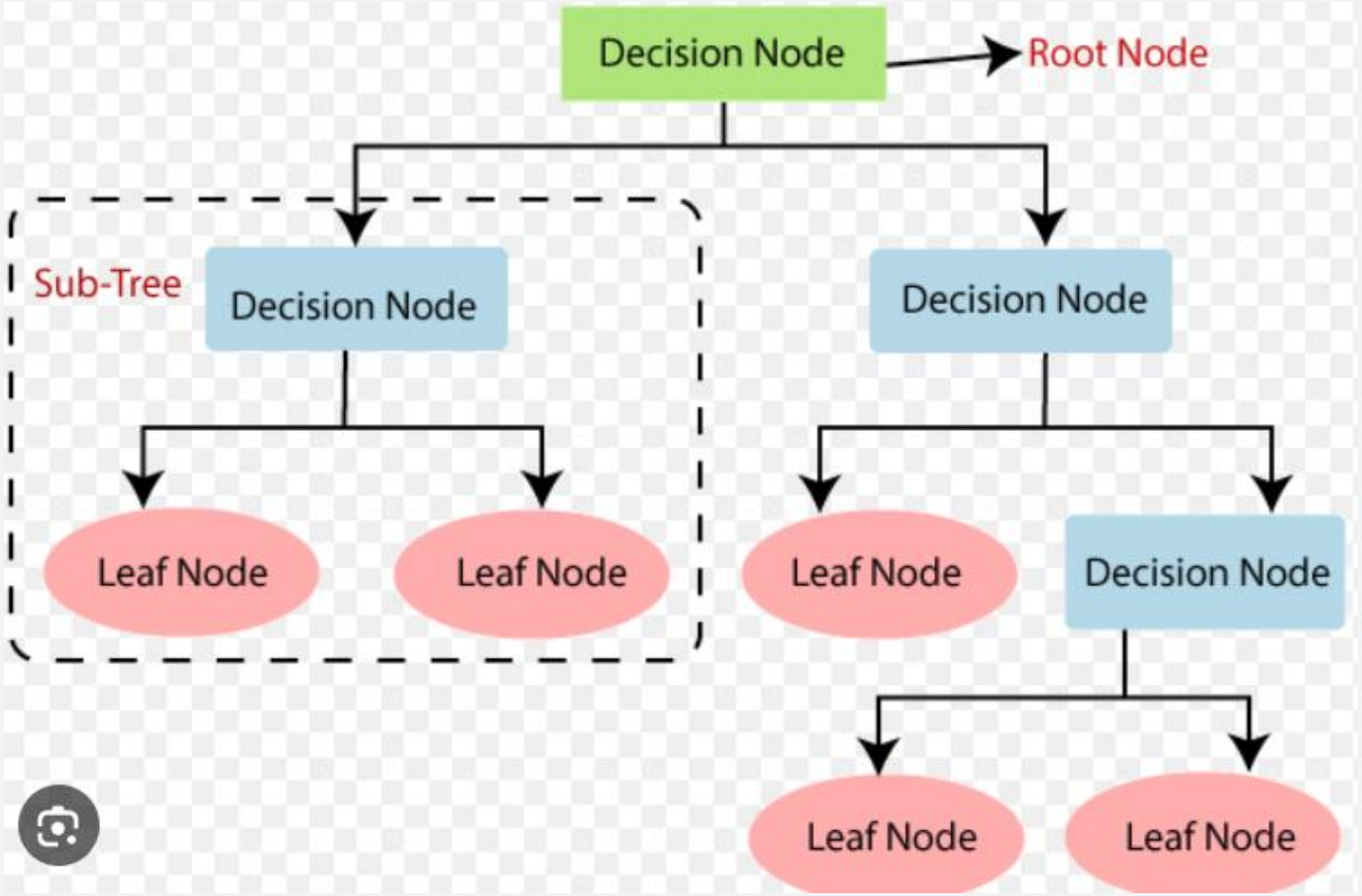
Mô hình hồi quy

Linear Regression và các biến thể

- Ordinary Least Squares Regression: $w = \min(\sum_{i=1}^{D}(y^{(i)} - Y^{(i)})^2$
- Lasso Regression: $w = \min(\sum_{i=1}^{D}(y^{(i)} - Y^{(i)})^2 + alpha * ||w||_1$
- Ridge Regression: $w = \min(\sum_{i=1}^{D}(y^{(i)} - Y^{(i)})^2 + alpha * ||w||_2^2$

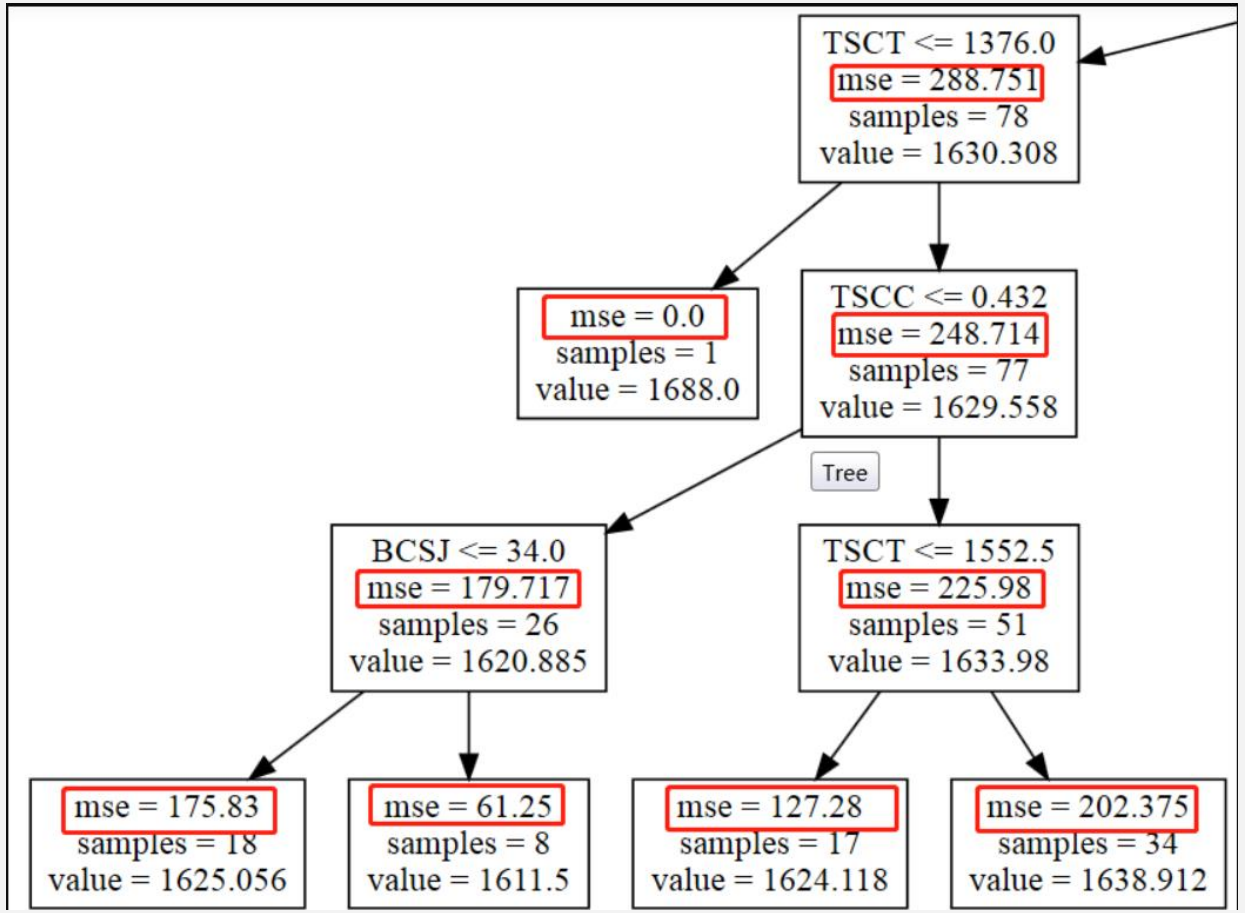
Mô hình hồi quy

Decision Tree Regression



Mô hình hồi quy

Decision Tree Regression



Mô hình hồi quy

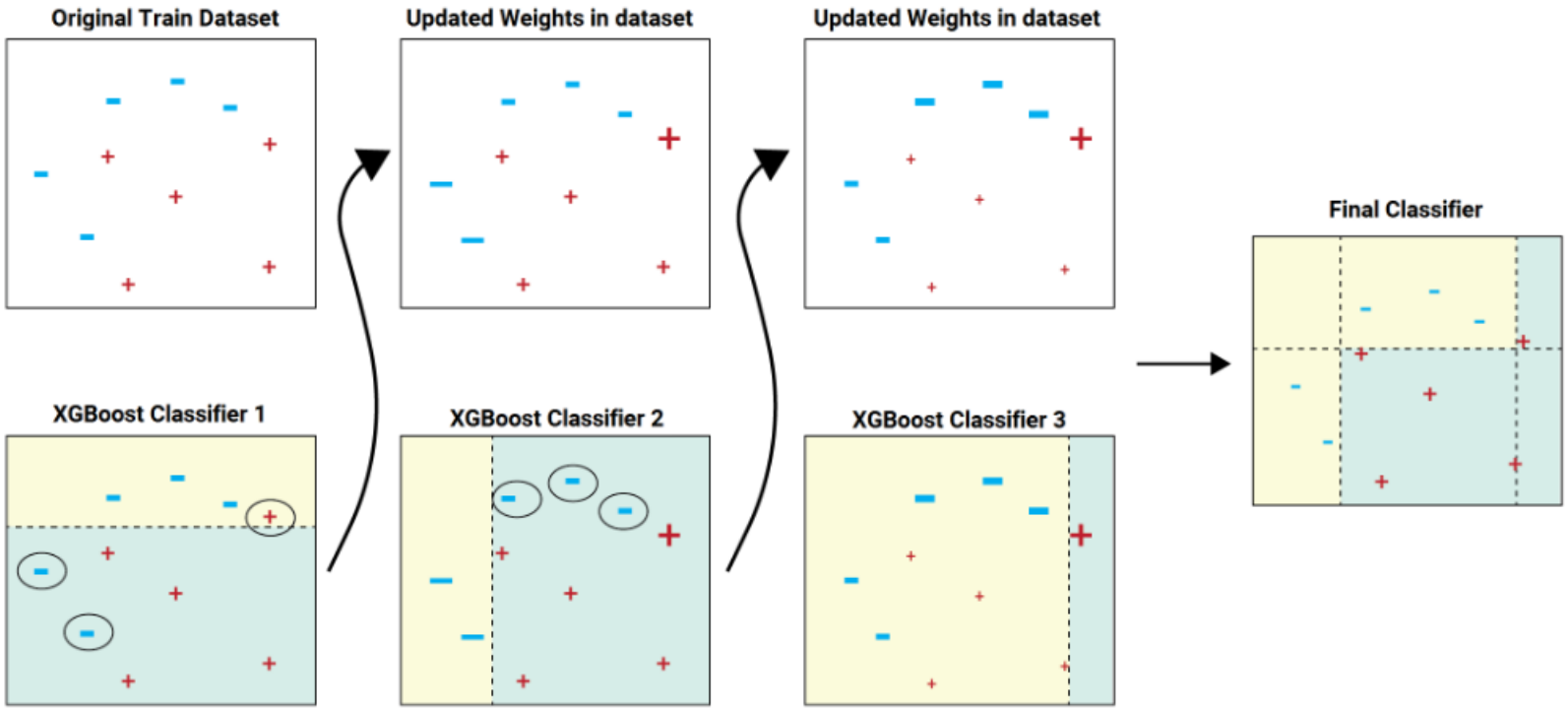
Xgboost Regression

- Có thể sử dụng cho cả phân loại và hồi quy
- Huấn luyện nhiều cây quyết định một cách tuần tự
- Mỗi cây quyết định là nông và được điều chỉnh với lỗi từ cây trước

➔ **Khi kết hợp sẽ tạo ra một mô hình có hiệu suất cao**

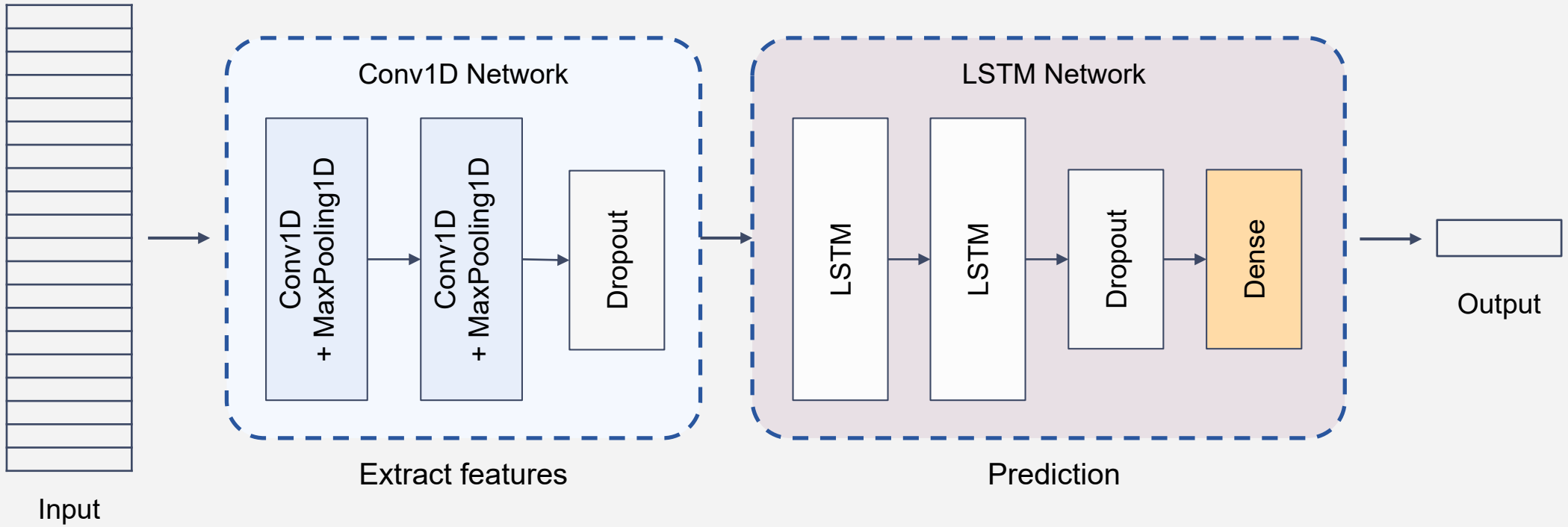
Mô hình hồi quy

Xgboost Regression



Neural Network

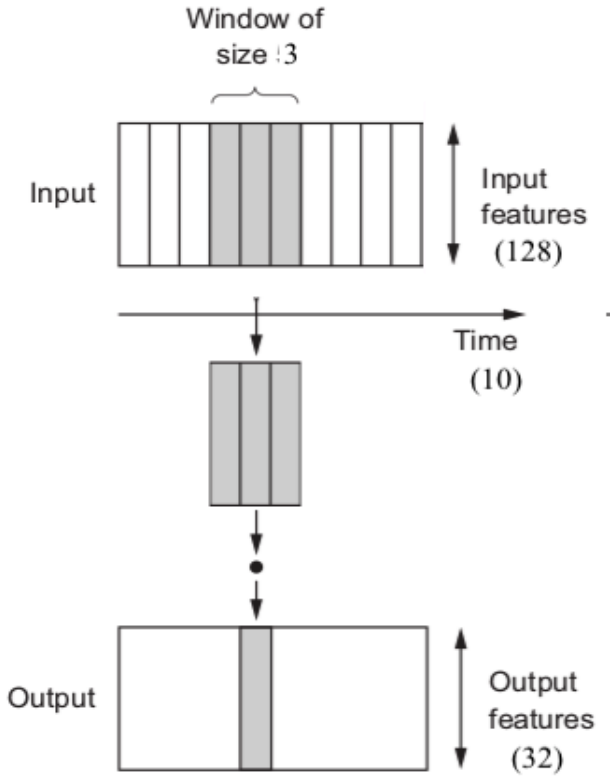
Mô hình Conv1D + LSTM



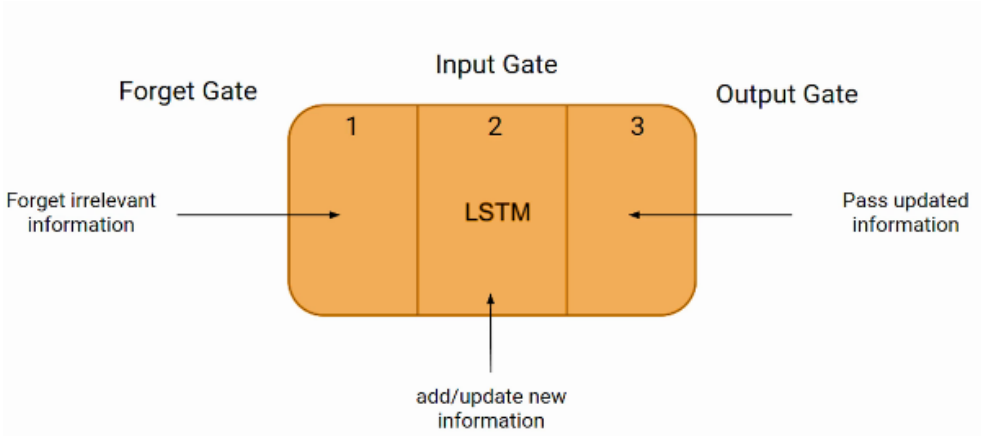
Neural Network

Mô hình Conv1D + LSTM

Conv1D



LSTM



Recommendation System

Tổng quan

- Một trong những ứng dụng phổ biến nhất của khoa học dữ liệu ngày nay.
- Được sử dụng để dự đoán "rating" hoặc "preference" mà người dùng sẽ dành cho một mặt hàng.
- Sinh viên giống như người dùng , môn học giống như sản phẩm, điểm môn học giống như đánh giá người dùng cho sản phẩm.

Recommendation System

Các bước chính

I. Tính toán độ tương đồng: Collaborative Filtering, Content Filtering, Hybrid Filtering

II. Dự đoán đánh giá dựa trên độ tương đồng:

Sau khi đã lấy được danh sách sinh viên tương đồng với sinh viên S. Ta chọn N sinh viên liên quan nhất. Sau đó với mỗi sinh viên, lấy điểm học tập P của môn học C và độ tương đồng U.

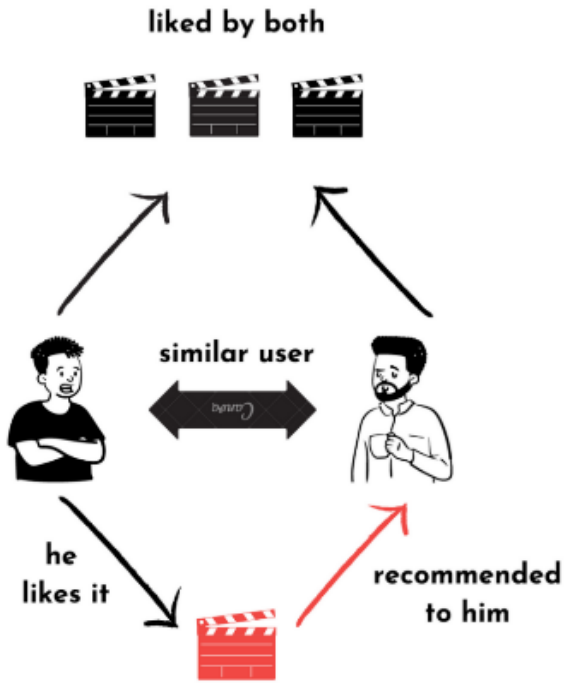
Khi đó, kết quả học tập môn học C của sinh viên S được dự đoán bằng:

$$P_C = (\sum_{c=1}^{N} P_c * U_c)/(\sum_{c=1}^{N} U_c)$$

Recommendation System

Collaborative Filtering

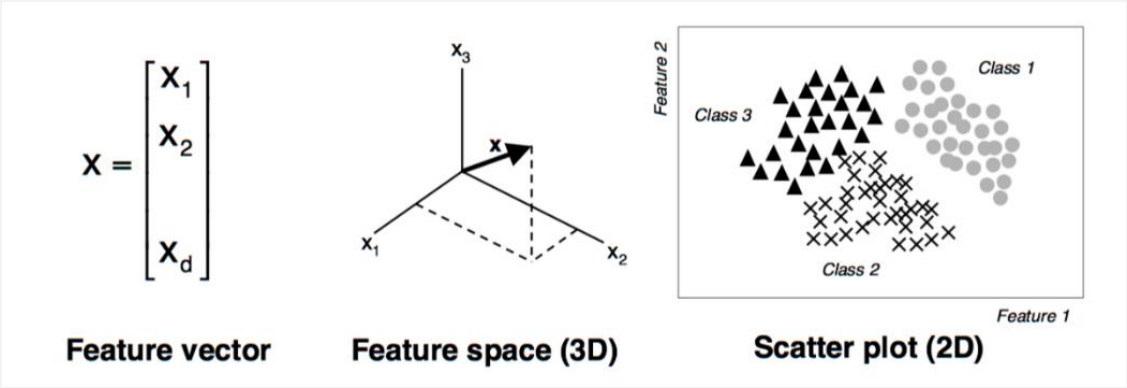
- Dự đoán kết quả những môn học mà sinh viên chưa học dựa trên **kết quả học tập** của các sinh viên tương tự

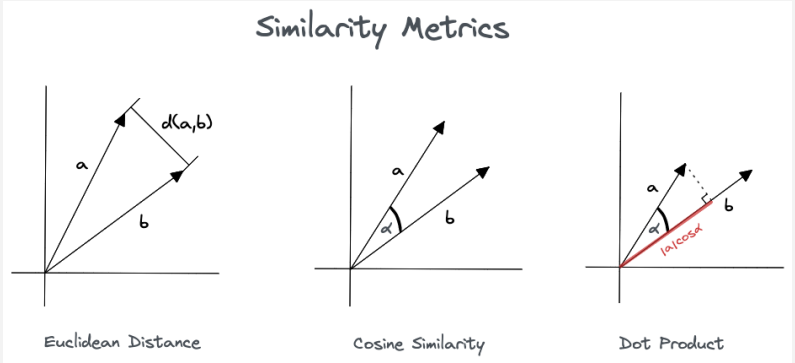


Recommendation System

Collaborative Filtering

- **Memory based:** sử dụng toàn bộ dữ liệu để dự đoán
  - User-based: dự đoán điểm môn C bằng cách chọn ra N sinh viên tương đồng nhất đã học môn đó
  - Item-based: dự đoán điểm sinh viên S bằng cách chọn ra N môn học tương đồng nhất sinh viên đã học

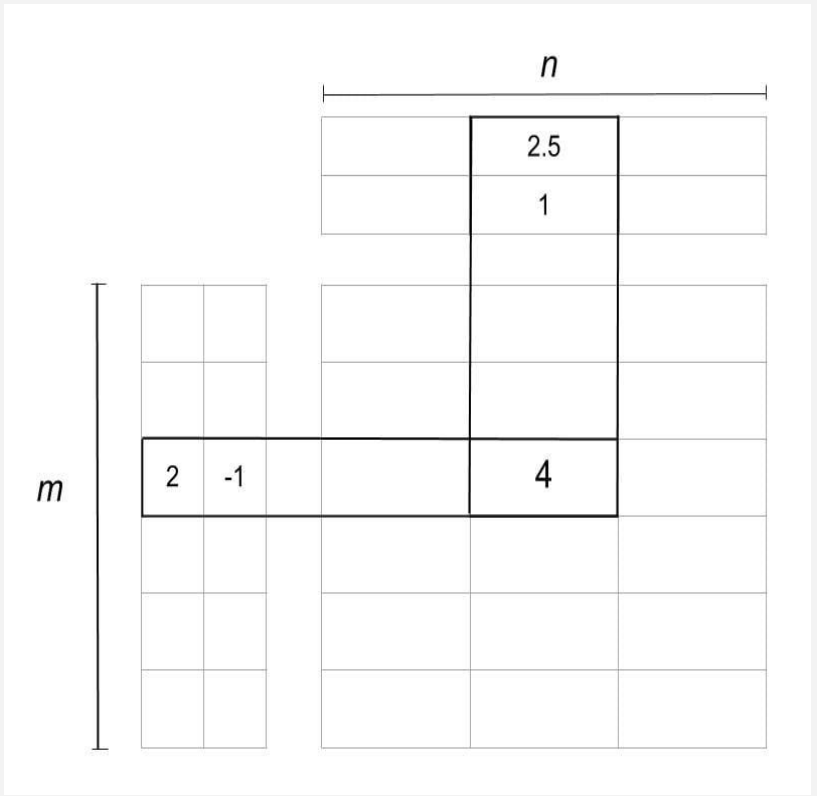




Recommendation System

Collaborative Filtering

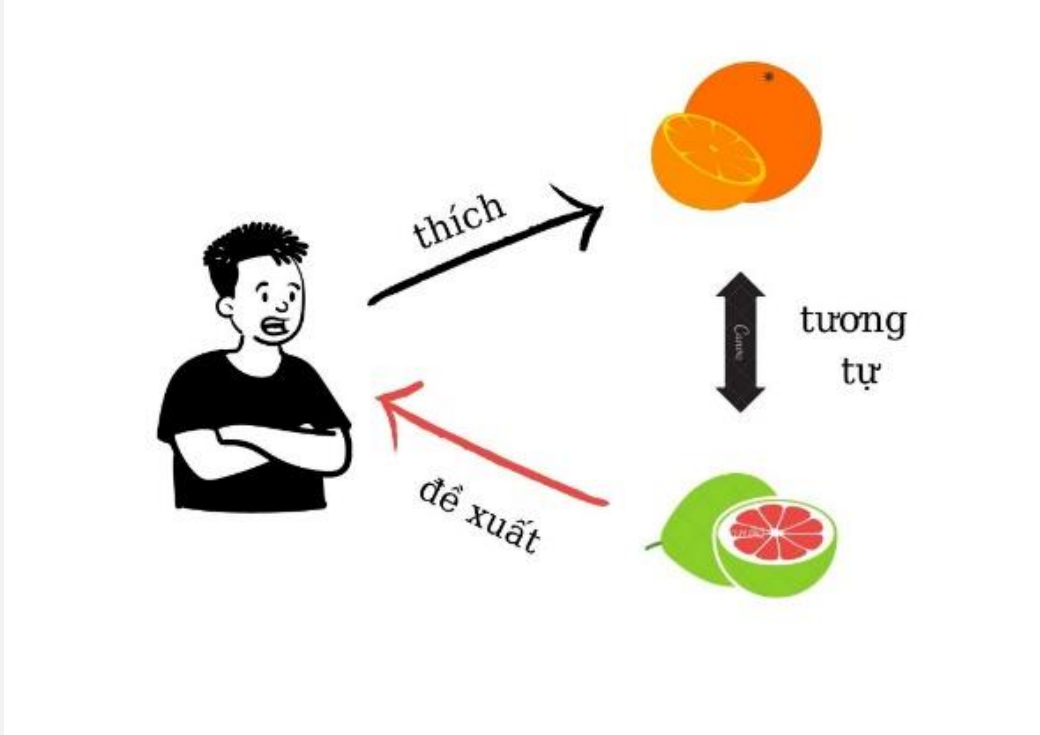
- **Model based:** làm giảm hoặc nén bộ dữ liệu ma trận thưa giữa sinh viên - môn học
  - Phân rã ma trận bằng thuật toán *SVD (Singular Value Decomposition)*



Recommendation System

Content Filtering

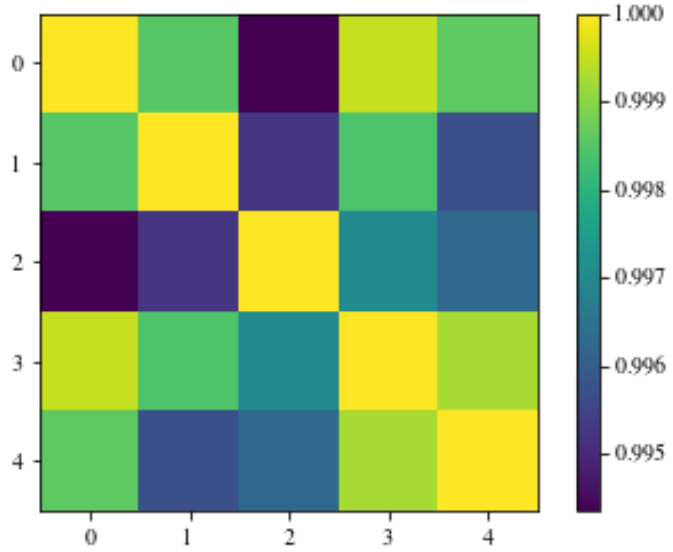
Dựa vào các **đặc tính của *sinh viên* và *môn học***, đề xuất ra các sinh viên và môn học có đặc trưng gần tương tự



Recommendation System

Content Filtering

- Đề xuất điểm dựa trên điểm các **môn học tương tự** mà sinh viên đó đã học
- Đề xuất điểm dựa trên điểm các **sinh viên tương tự** đã học qua môn học đó



Thực nghiệm

Root Mean Squared Error

STT	Phương pháp	Test RMSE
1	Collaborative Filtering - Model based	1.7200
2	Collaborative Filtering - Memory based	1.7258
3	Conv1D_LSTM	1.76854
4	Xgboost Regression	1.79939
5	Lasso Regression	1.88168
6	Ordinary Least Squares Regression	1.88256
7	Ridge Regression	1.88256
8	Content Filtering – Student based	1.8851
9	Decision Tree Regression	1.88814
10	Content Filtering - Course based	2.2207

Kết luận và hướng phát triển

- Hướng phát triển:
  - Tối ưu hóa các siêu tham số
  - Thử nghiệm thêm các phương pháp khác

-THE END-

THANKS FOR LISTENING

-Group 3-