BÁO CÁO ĐỒ ÁN

Môn: Mạng xã hội - ThS. Nguyễn Thị Anh Thư

Đề tài: **Dự đoán chất lượng khóa học - MOOCCubeX**

Thành viên nhóm 3:

- Hoàng Ngọc Quý (NT)
- Trần Khiết Tường
- Nguyễn Tuấn Khang
- Nguyễn Trường Huy
- Nguyễn Đức Thịnh

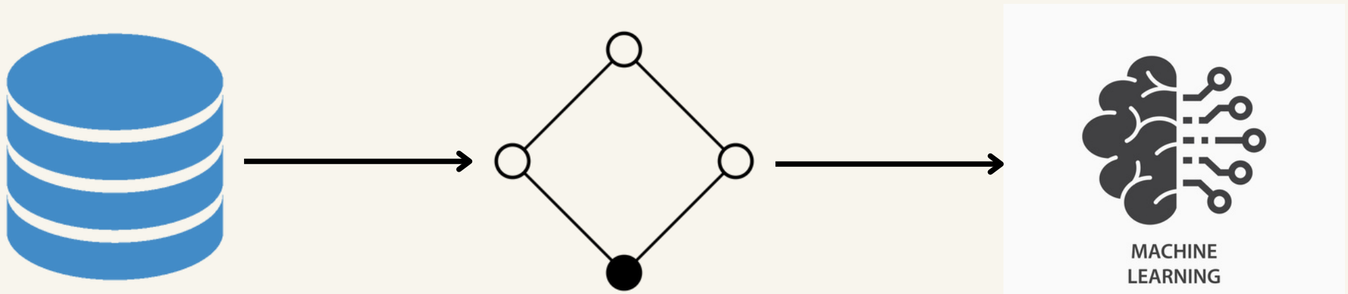
1 - GIỚI THIỆU ĐỀ TÀI

1 - GIỚI THIỆU ĐỀ TÀI

Đề tài : Dự đoán chất lượng khóa học qua dataset MOOCCubeX

Mục tiêu đề tài:

- Áp dụng xây dựng mô hình mạng vào bài toán dự đoán chất lượng khóa học để model đạt được kết quả tốt hơn



Dataset MOOCCubeX Xây dựng mô hình mạng Huấn luyện mô hình

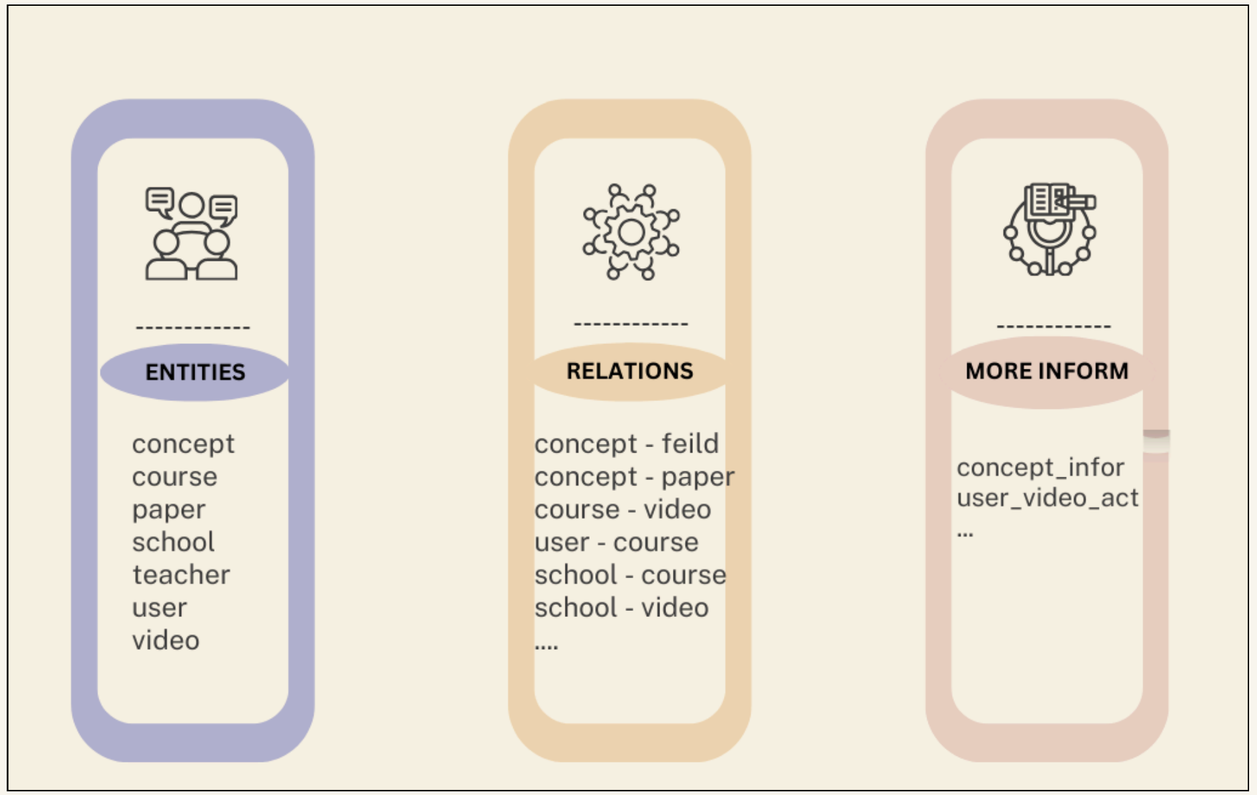
2 - DATASET

2 - DATASET

Nguồn: Bộ dữ liệu MOOCCubeX được xây dựng bởi The Knowledge Engineering Group của trường Đại học Thanh Hoa và được hỗ trợ bởi XuetangX.

Thông tin miêu tả:

- Khóa học 706
- Video 38.181
- User 199.199
- Khái niệm 114.563



2 - DATASET - EDA

Thăm dò dữ liệu:

```
user = get_data_csv("/content/mooc/MOOCCube/entities/user.json", "/content/user.csv")

print(user.head())
```

```
           id name                                        course_order  \
0   U_7001215  李寒锋  ['C_course-v1:TsinghuaX+00740043_2x_2015_T2+sp...
1  U_10402446  五元香  ['C_course-v1:TsinghuaX+00510888X+2019_T1', 'C...
2  U_10359065   魏珊  ['C_course-v1:TsinghuaX+00612642X+sp', 'C_cour...
3   U_7423998  郭海滨  ['C_course-v1:TsinghuaX+30240184_2X+sp', 'C_co...
4    U_545306  李其艳  ['C_course-v1:TsinghuaX+20430064_2X+sp', 'C_co...

                                     enroll_time
0  ['2017-05-01 11:07:53', '2017-05-17 10:07:17',...
1     ['2019-06-14 08:50:04', '2019-01-04 20:36:07']
2     ['2019-01-18 21:19:56', '2019-01-14 21:54:54']
3  ['2017-08-16 10:38:11', '2018-07-01 18:24:24',...
4  ['2018-09-05 15:40:40', '2019-02-28 10:08:49',...
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 199199 entries, 0 to 199198
Data columns (total 4 columns):
 #   Column        Non-Null Count   Dtype 
---  ------        --------------   ----- 
 0   id            199199 non-null  object
 1   name          199199 non-null  object
 2   course_order  199199 non-null  object
 3   enroll_time   199199 non-null  object
dtypes: object(4)
memory usage: 6.1+ MB
None
```

User

2 - DATASET - EDA

Thăm dò dữ liệu:

```
school = get_data_csv("/content/mooc/MOOCCube/entities/school.json", "/content/school.csv")

print(school.head())
```

```
        id     name                                                about
0    S_BNU  北京师范大学  北京师范大学 (Beijing Normal University) 简称"北师大", 由中华人民...
1    S_UQx   昆士兰大学  昆士兰大学 (The University of Queensland) , 简称昆大, 世界50强...
2  S_UESTC  电子科技大学  电子科技大学 (University of Electronic Science and Te...
3    S_BIT  北京理工大学  北京理工大学 (Beijing Institute of Technology) 简称北理工, 由...
4    S_JNU    暨南大学  暨南大学 (Jinan University) , 简称""暨大 (JNU) , 位于广东省广州市, 是中...
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 208 entries, 0 to 207
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   id      208 non-null    object
 1   name    194 non-null    object
 2   about   194 non-null    object
dtypes: object(3)
memory usage: 5.0+ KB
None
```

School

2 - DATASET - EDA

Thăm dò dữ liệu:

```
course = get_data_csv("/content/mooc/MOOCCube/entities/course.json", "/content/course.csv")

print(course.head())
```

```
                                          id                  name  \
0          C_course-v1:McGillX+ATOC185x+2015_T1          自然灾害（自主模式）
1     C_course-v1:TsinghuaX+THESIS2015X+2015_T1  2015年清华大学研究生学位论文答辩（二）
2  C_course-v1:TsinghuaX+THESIS2014_1X_tvs_2014_  2014年清华大学研究生学位论文答辩（一）
3   C_course-v1:TsinghuaX+THESIS2015X_tvs2015_T1  2015年清华大学研究生学位论文答辩（一）
4              C_course-v1:TsinghuaX+00600242+sp       文物精品与文化中国（自主模式）

  prerequisites                                              about  \
0            无  <p>地球上没有一处地方不发生自然灾害，当我们以科学的眼光看待这些自然灾害的原因和本质时，我...
1        无先修要求  <p>学位论文答辩环节是研究生培养的重要环节，为了充分发挥该环节的育人作用，搭建学术交流的平...
2        无先修要求  <p>学位论文答辩环节是研究生培养的重要环节，为了充分发挥该环节的育人作用，搭建学术交流的平...
3        无先修要求  <p>学位论文答辩环节是研究生培养的重要环节，为了充分发挥该环节的育人作用，搭建学术交流的平...
4            无  <p>中国考古学是以往100年中发展最为迅速的领域之一，大批珍贵文物的出土，不断刷新人们对文...

                                  core_id  \
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 706 entries, 0 to 705
Data columns (total 8 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   id             706 non-null    object
 1   name           706 non-null    object
 2   prerequisites  702 non-null    object
 3   about          706 non-null    object
 4   core_id        706 non-null    object
 5   video_order    706 non-null    object
 6   display_name   706 non-null    object
 7   chapter        706 non-null    object
dtypes: object(8)
memory usage: 44.2+ KB
None
```

Course

2 - DATASET - EDA

Thăm dò dữ liệu:

```
teacher = get_data_csv("/content/mooc/MOOCCube/entities/teacher.json", "/content/teacher.csv")

print(teacher.head())
```

```
      id name                                                about
0  T_方维奇  方维奇  方维奇，男，陕西工业职业技术学院教师，讲师。主持并参与院级科研项目3项，发表教科研论文13篇...
1  T_范茂魁  范茂魁  范茂魁（1979-）男，教授，硕士学位，教龄17年，主要从事《消防燃烧学》、《化学基础》、《...
2  T_连小琨  连小琨  连小琨，男，1955年生于重庆。清华大学教授，清华大学机械工程学位分委会副主席。中国汽车工程...
3  T_张德顺  张德顺  张德顺,1964年生,同济大学建筑与城市规划学院高密度人居环境生态与节能 教育部重点实验室教...
4   T_王红   王红   王红，硕士，教授，硕士生导师。长期从事害生虫学教学 与科研工作。教学方面，独立系统地担任了本...
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1738 entries, 0 to 1737
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   id      1738 non-null   object
 1   name    1738 non-null   object
 2   about   1738 non-null   object
dtypes: object(3)
memory usage: 40.9+ KB
None
```

Teacher

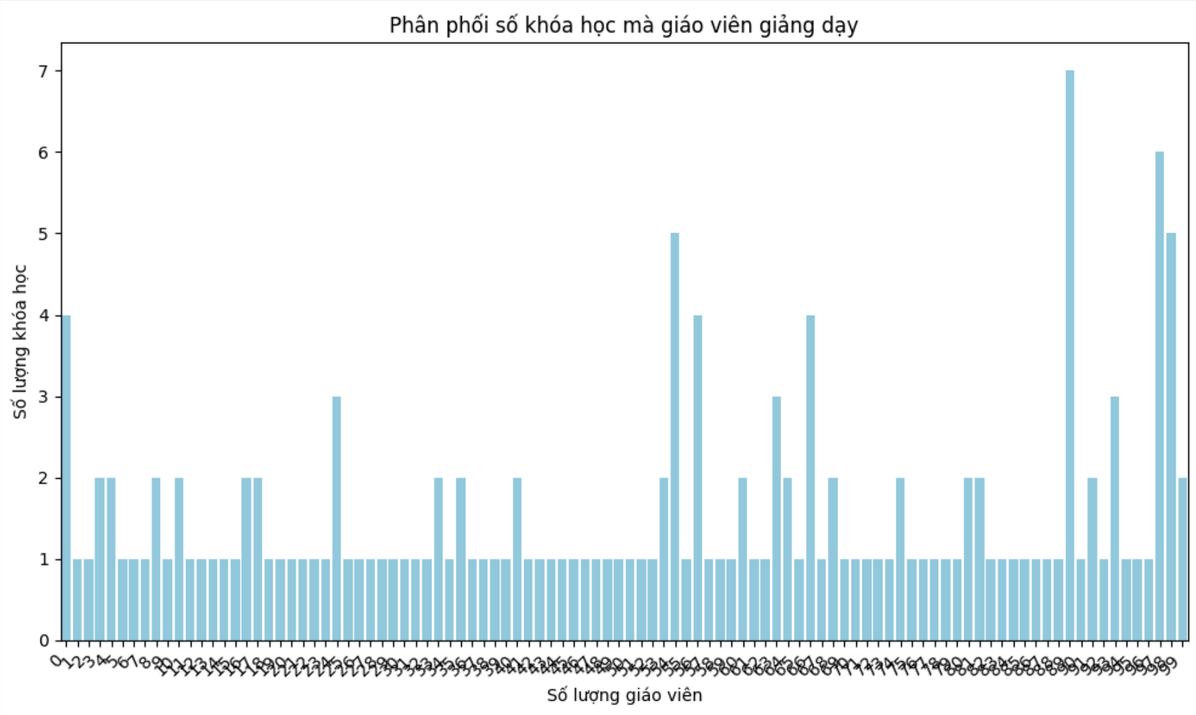
2 - DATASET - EDA

Thăm dò dữ liệu:

```
teacher_course = read_json_relation("/content/mooc/MOOCCube/relations/teacher-course.json", "/co

print(teacher_course.head())
```

```
  teacher                                    course
0   T_方维蓉           C_course-v1:SPI+20170828001x+sp
1   T_方维蓉      C_course-v1:SXPI+20170828001x+2019_T1
2   T_郑茂魁      C_course-v1:PSFFC+2018102405X+2018_T2
3   T_连小琨  C_course-v1:TsinghuaX+70150104_2X+2019_T1
4   T_连小琨    C_course-v1:TsinghuaX+70150104X+2019_T1
```



Course_Teacher

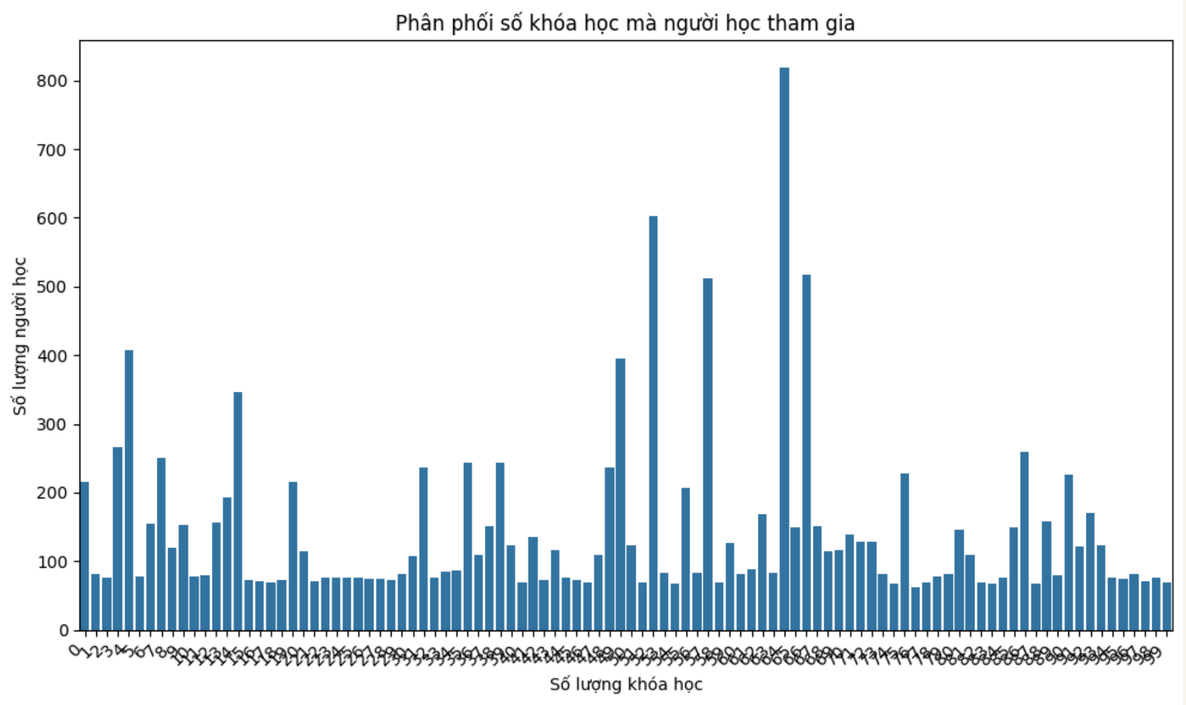
2 - DATASET - EDA

Thăm dò dữ liệu:

```
user_course = read_json_relation("/content/mooc/MOOCCube/relations/user-course.json", "/content/user-course.csv", ["user","course"])

print(user_course.head())
```

```
      user                                          course
0  U_7001215  C_course-v1:TsinghuaX+00740043_2x_2015_T2+sp
1  U_7001215             C_course-v1:TsinghuaX+30240184+sp
2  U_7001215    C_course-v1:TsinghuaX+00740043X_2015_T2+sp
3  U_7001215     C_course-v1:TsinghuaX+10421094X_2015_2+sp
4  U_7001215          C_course-v1:TsinghuaX+30240184_2X+sp
```



Course_User

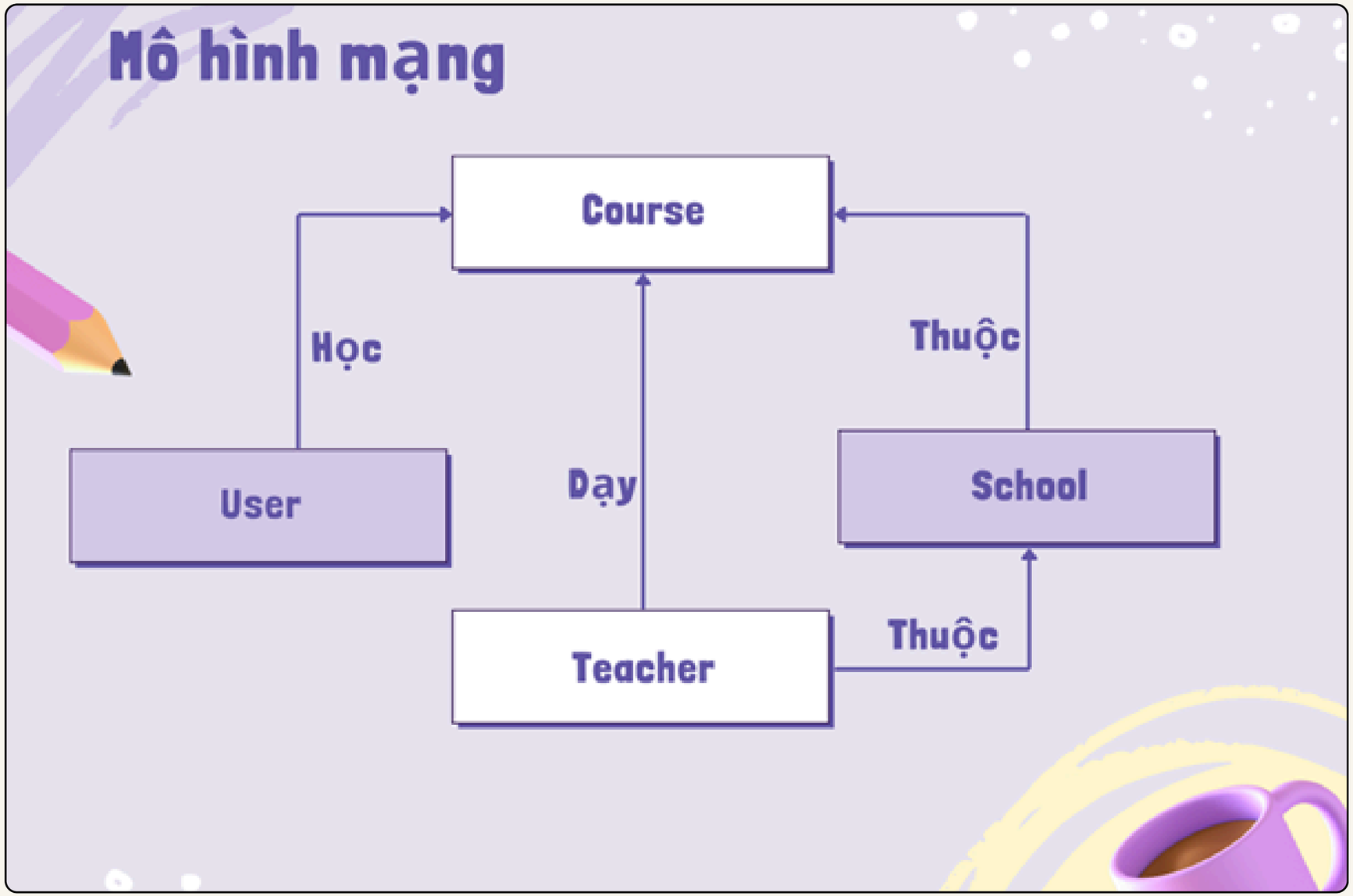
2 - DATASET - Data Preprocessing

Tiền xử lý dữ liệu: Các dữ liệu trong dataset tương đối sạch như đã EDA trước đó các dữ liệu thường rất ít xuất hiện các giá trị null, NaN hay Outliers. Nếu xảy ra trường hợp sẽ xử lí theo 2 cách:

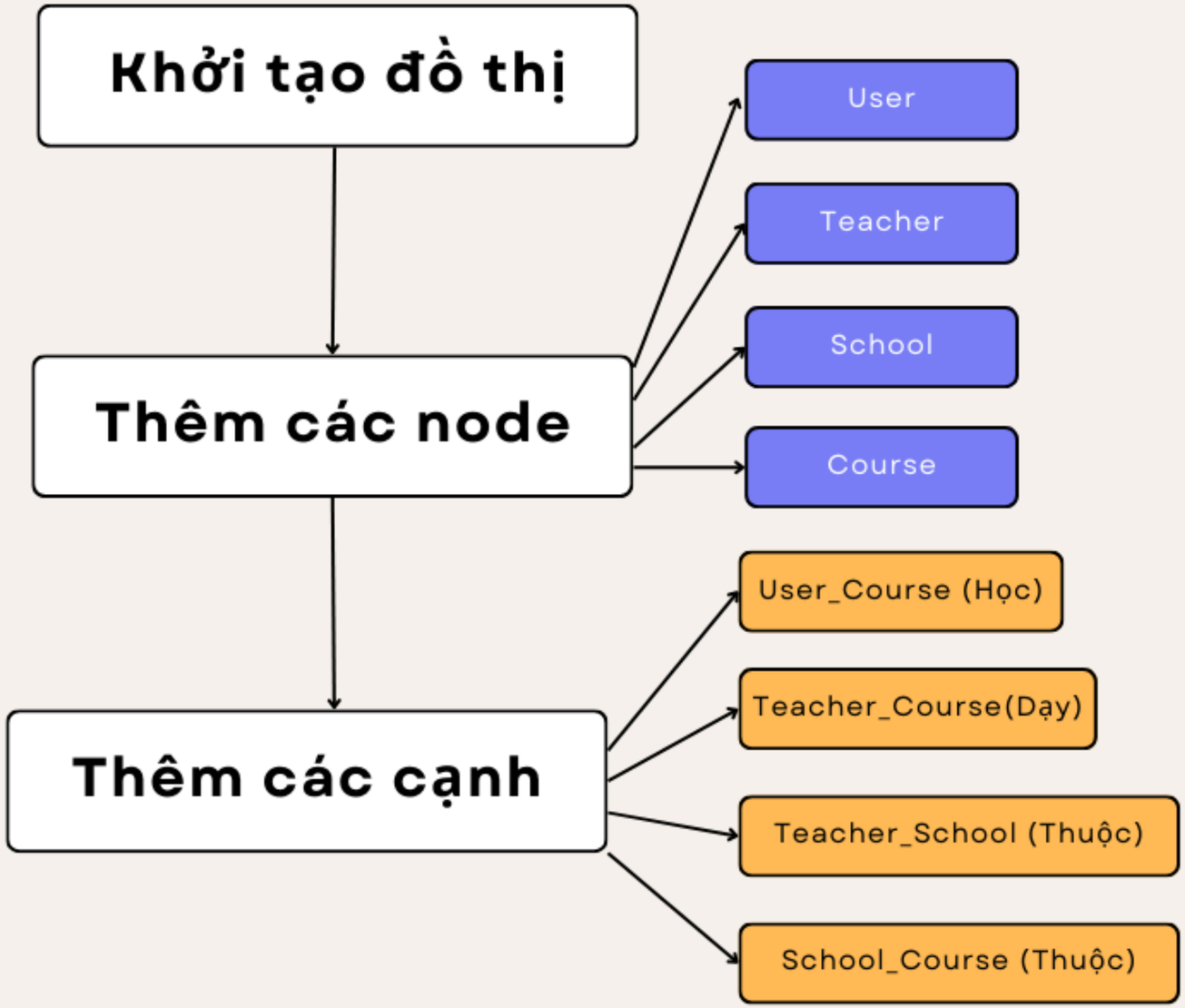
+ Xóa thể hiện chứa các giá trị null, NaN, Outliers
+ Inner Join các Dataset với dữ liệu có trùng ID

3- XÂY DỰNG MÔ HÌNH MẠNG

3 - MÔ HÌNH MẠNG



3 - MÔ HÌNH MẠNG



4 - THUẬT TOÁN

4 - THUẬT TOÁN

+ Sử dụng Node2Vec để embedding từ dữ liệu mô hình mạng.

+ Các input đầu vào: course.json, user.json, school.json, teacher.json, user - problem.json.

+ Tiêu chí đánh giá: tỷ lệ điểm số của các user đối với exercise đạt được so với điểm tối đa .

+ Mô hình sử dụng: SVM, RandomForestClassifier, GradientBoostingClassifier.

4 - THUẬT TOÁN

+ Kết quả

	SVM	Random Forest Classifier	Gradient Boosting Classifier
accuracy	0.9	0.93	0.87
precision	1.0	1.0	1.0
recall	0.9	0.93	0.86
F1 - score	0.95	0.96	0.93

4 - THUẬT TOÁN

+ Nhận xét

Random Forest: Hiệu suất toàn diện tốt nhất, cân bằng giữa độ chính xác và khả năng nhận diện.

SVM: Độ chính xác cao khi phân loại mẫu âm nhưng có thể bỏ sót một số mẫu dương.

Gradient Boosting: Hiệu suất thấp nhất, đặc biệt là khả năng nhận diện mẫu dương.

-> Kết luận: Random Forest là mô hình được khuyến nghị sử dụng.

5 - KẾT LUẬN

5 - KẾT LUẬN

Kết luận:

Nhóm đã thành công trong việc áp dụng kiến thức cơ bản về mô hình mạng, giúp nâng cao hiệu quả của mô hình học máy và cải thiện kết quả phân tích.

Hướng phát triển:

Cần huấn luyện thêm nhiều mô hình học máy để đưa ra đánh giá và so sánh ưu nhược điểm của từng thuật toán. Đồng thời, khai phá thêm các thông tin tiềm ẩn trong mô hình mạng và áp dụng các chỉ số trung tâm, phân cụm để cải thiện chất lượng dữ liệu. Ngoài ra, cần phát triển thêm một ứng dụng demo để trình bày kết quả thực tế.

THANKS