Một số chủ đề đồ án:

1. Phân tích và so sánh các phương pháp học máy cho 1 bài toán cụ thể
2. Phân tích và so sánh sự hiệu quả của các bước tiền xử lý trong một bài toán cụ thể
3. Tìm siêu tham số tối ưu (được tìm qua các phương pháp optimization) và so sánh tính hiệu quả của bộ siêu tham số này với các bộ siêu tham số khác
4. Phân tích ảnh hưởng của các yếu tố lên kết quả của một dữ liệu nào đó, sau đó đưa ra mô hình regression tối ưu

Một số data có thể dùng:

1. Data cho bài toán regression, time series:
   https://archive.ics.uci.edu/ml/datasets.php?format=&task=&att=&area=&numAtt=&numIns=&type=ts&sort=taskUp&view=table

   https://data.world/datasets/time-series

2. Data tự thu thập về giá nhà, giá cổ phiếu, giá phòng khách sạn, lưu lượng giao thông, chất lượng nước, không khí, sản lượng nông nghiệp, số khách du lịch,…
   https://www.gso.gov.vn/
3. Data tổng hợp từ nhiều nguồn. Ví dụ lấy data về sản lượng, giá lúa ở 1 nguồn và data về thời tiết, giá phân bón và các yếu tố khác ở các nguồn khác.

Cách chọn đồ án để có thể phát triển thêm:

1. Cần phải phát biểu được tính ứng dụng, sự cần thiết và đóng góp của đồ án
2. Phát biểu được tính mới (tương đối) của đồ án (mới về cách xử lý, mới về data, mới vì chưa ai dùng cách xử lý này trên data này,etc.,)
3. **QUAN TRỌNG**: nên chọn các bài toán mà mình khảo sát thấy hiện tại người ta hay giải quyết theo cách đơn giản. Ví dụ một số dữ liệu timeseries hiện tại chỉ được áp dụng các mô hình regression và các phương pháp ML đơn giản để dự đoán. Do đó nếu ta áp dụng các mô hình timeseries, đặc biệt áp dụng mô hình đa biến thì khả năng sẽ cải thiện được độ chính xác. Khi so sánh các mô hình thì kết quả sẽ thấy rõ.
4. Chọn dữ liệu việt nam hay quốc tế? Nếu chọn dữ liệu quốc tế thì dữ liệu thường có sẵn, có thu thập thêm cũng dễ. Tuy nhiên cần áp dụng các phương pháp phức tạp hơn cho dữ liệu này để có tính mới. Nếu chọn dữ liệu Việt Nam thì thu thập dữ liệu sẽ mất công hơn. Tuy nhiên do các dữ liệu này ít được xử lý và các phương pháp xử lý hiện có thường đơn giản nên ta sẽ dễ dàng áp dụng các phương pháp phức tạp hơn một tí để có được tính mới. Ví dụ hiện tại có nhiều bài về dự đoán các data trong nước nhưng sử dụng các phương pháp đơn giản được đăng ở các tạp chí trong nước. Do đó nếu ta cũng dùng các data này và áp dụng phương pháp phức tạp hơn là đã có thể phát biểu được tính mới.
   https://drive.google.com/drive/folders/1N4DJeDJzpiC6AHKEItylatdwoe4m3Es6?usp=sharing