Project

1 Mô tả

Đồ án này sẽ dựa trên cuộc thi Prediction Interval Competition II - House Price. Nhóm của bạn sẽ xây dựng một mô hình dự đoán giá nhà dựa trên các đặc trưng đầu vào. Mô hình cần dự đoán khoảng giá có độ tin cậy cao, thay vì chỉ dự đoán giá trị trung bình, nhằm phục vụ cho các ứng dụng thực tế như tư vấn tài chính, bảo hiểm, và định giá tài sản.

Mục tiêu của đồ án là giúp các bạn thực hành các kỹ thuật phân tích và xử lý dữ liệu bảng (tabular data), đồng thời hiểu thêm về bài toán dự đoán có ràng buộc khoảng tin cậy (prediction interval regression).

Các công việc yêu cầu cụ thể được trình bày dưới đây.

2 Công việc cụ thể

2.1 EDA và xử lý dữ liệu (Exploratory Data Analysis & Data Preprocessing)

Ở bước này, nhóm cần thực hiện việc phân tích sơ bộ để hiểu rõ hơn về tập dữ liệu cũng như tiến hành các bước xử lý dữ liệu ban đầu nhằm chuẩn bị cho việc huấn luyện mô hình. Một số gợi ý:

- EDA: Thống kê mô tả các thuộc tính, Phân tích sự phân bố của giá nhà,....
- Xử lý dữ liệu: loại bỏ ngoại lai, chuẩn hóa các giá trị số,...

2.2 Huấn luyện mô hình dự đoán khoảng giá

Sinh viên cần huấn luyện một mô hình dự đoán khoảng giá nhà (gồm giới hạn dưới và giới hạn trên) trên dữ liệu huấn luyện được cung cấp. Kết quả đánh giá dựa trên điểm số đánh giá trên hệ thống Kaggle. Các nhóm phải submit kết quả lên hệ thống để kiểm tra hiệu năng.

3 Các nội dung cần nộp

1. **Source code**: gồm các folder con `Q1, Q2` tương ứng với các yêu cầu 2.1, 2.2. Mỗi phần cần có file `README.md` hướng dẫn cách chạy, đảm bảo chạy được trên nền tảng Google Colab.
2. **Báo cáo**: gồm các nội dung sau:
   (a) Bìa, mục lục
   (b) Tự đánh giá mức độ hoàn thành, phân công và đóng góp của từng thành viên
   (c) Nội dung chính chia thành các phần:
      i. EDA: các biểu đồ, bảng phân tích, nhận xét
      ii. Xử lý dữ liệu: các phương pháp xử lý được sử dụng, mô tả
      iii. Mô hình dự đoán: mô tả mô hình, hiệu năng, ưu nhược điểm, kết quả trên tập kiểm thử, kết quả, nhận xét

3. **Video thuyết trình ngắn (10–20 phút)**: Trình bày slide về quá trình và kết quả đồ án. Nếu video lớn, có thể upload lên YouTube/Drive và chèn link vào báo cáo.
4. **Slide trình bày**

4 Lưu ý

- Đồ án sẽ được đánh giá theo khối lượng công việc chia theo số lượng thành viên. Báo cáo chiếm 50% điểm tổng, cần trình bày rõ ràng, có chiều sâu.
- Mỗi nhóm thành viên tương tự như nhóm trên lớp lý thuyết. Khi tham gia cuộc thi Kaggle, bắt đầu tên đội theo cú pháp: `FIT-HCMUS-<GroupName>`.
- Đạo văn hoặc sao chép sẽ bị xử lý nghiêm và có thể dẫn đến 0 điểm toàn môn. Trong trường hợp nghi ngờ, GV có thể yêu cầu vấn đáp.