HOME WORK 6

Bài tập 1

1. Các phương pháp/ công thức đánh giá độ chính xác của thuật toán/ mô hình dự báo.
2. **Phân loại (Classification)**

- True Positive (TP)

- True Negative (TN)

- False Positive (FP)

- False Negative (FN)

**Độ chính xác (Precision):** Tỷ lệ các trường hợp dự đoán dương tính chính xác so với tổng số trường hợp được dự đoán dương tính.

[Precision = TP / (TP + FP)]

**Độ nhớ lại (Recall):** Tỷ lệ các trường hợp dương tính thực tế được dự đoán chính xác so với tổng số trường hợp dương tính thực tế.

[Recall = TP / (TP + FN)]

**Điểm F1 (F1-Score):** Trung bình điều hòa của Precision và Recall.

[F1 = 2 \* (Precision \* Recall) / (Precision + Recall)]

**Đường cong ROC (ROC Curve):** Biểu đồ thể hiện tỷ lệ True Positive (TP) so với False Positive (FP) tại các ngưỡng phân loại khác nhau.

**Diện tích dưới đường cong ROC (AUC):** Đo lường khả năng phân biệt giữa các lớp của mô hình. Giá trị AUC càng cao, khả năng phân biệt càng tốt.

**2. Hồi quy (Regression)**

**Sai số tuyệt đối trung bình (Mean Absolute Error - MAE):** Trung bình cộng của các sai số tuyệt đối giữa giá trị thực và giá trị dự báo.

[MAE = 1/n \* Σ(i=1 đến n) |yi - ŷi|]

**Sai số bình phương trung bình (Mean Squared Error - MSE):** Trung bình cộng của bình phương các sai số giữa giá trị thực và giá trị dự báo.

[MSE = 1/n \* Σ(i=1 đến n) (yi - ŷi)^2]

**Hệ số xác định (R-squared - R²):** Tỷ lệ phần trăm của phương sai trong biến phụ thuộc được giải thích bởi mô hình hồi quy. [R^2 = 1 - (Σ(i=1 đến n) (yi - ŷi)^2) / Σ(i=1 đến n) (yi - ȳ)^2]

Trong đó:

* ​ là giá trị thực tế.
* ​ là giá trị dự báo.
* là số lượng mẫu.

Giả sử chúng ta có một mô hình dự báo giá nhà dựa trên một số đặc điểm của nhà. Chúng ta có dữ liệu thực tế và dữ liệu dự báo như sau:

|  |  |  |
| --- | --- | --- |
| **Nhà** | **Giá trị thực tế ($)** | **Giá trị dự báo ($)** |
| 1 | 300,000 | 310,000 |
| 2 | 450,000 | 430,000 |
| 3 | 500,000 | 520,000 |
| 4 | 350,000 | 340,000 |
| 5 | 600,000 | 610,000 |

**Bước 1:** Tính toán sai số tuyệt đối cho từng cặp giá trị thực tế và giá trị dự báo.

**Bước 2:** Tính tổng của các sai số tuyệt đối.

**Bước 3:** Chia tổng sai số tuyệt đối cho số lượng mẫu để tính MAE.

Kết quả

MAE cho mô hình dự báo này là 14,000 đô la. Điều này có nghĩa là, trung bình, dự báo của mô hình sai lệch 14,000 đô la so với giá trị thực tế.

**Bước 1:** Tính toán sai số bình phương cho từng cặp giá trị thực tế và giá trị dự báo.

**Bước 2:** Tính tổng của các sai số bình phương.

**Bước 3:** Chia tổng sai số bình phương cho số lượng mẫu để tính MSE.

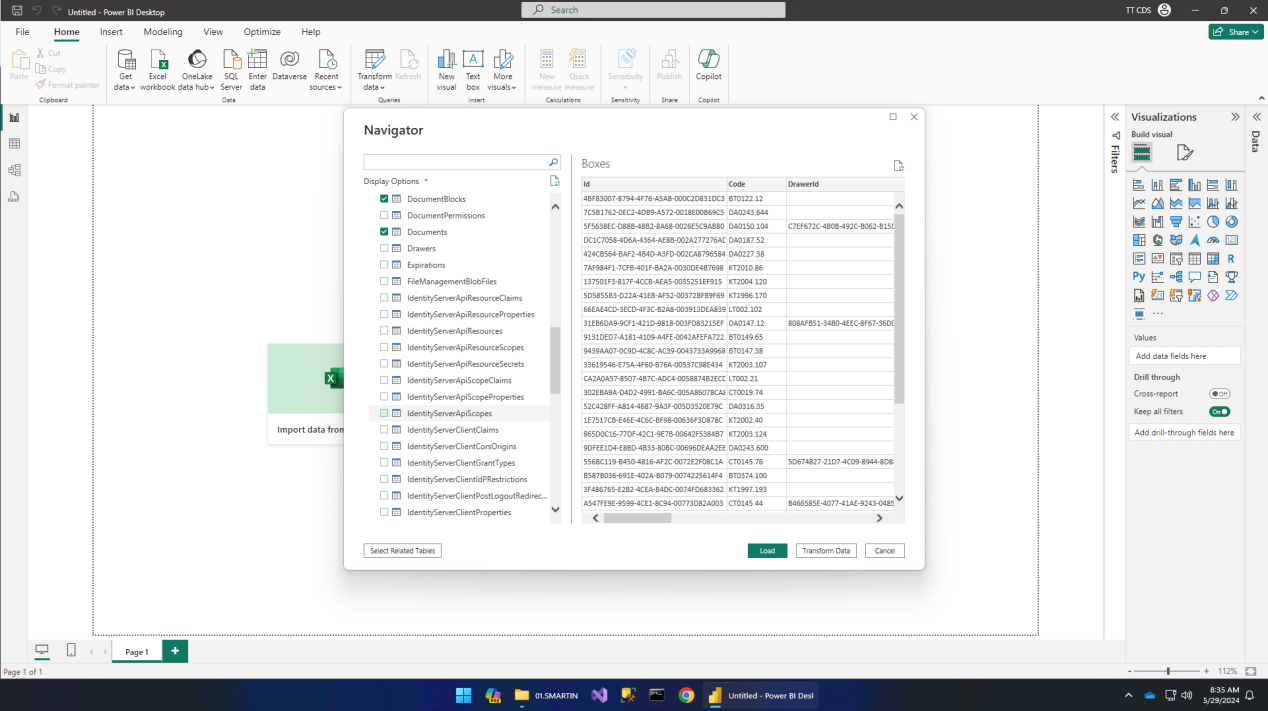
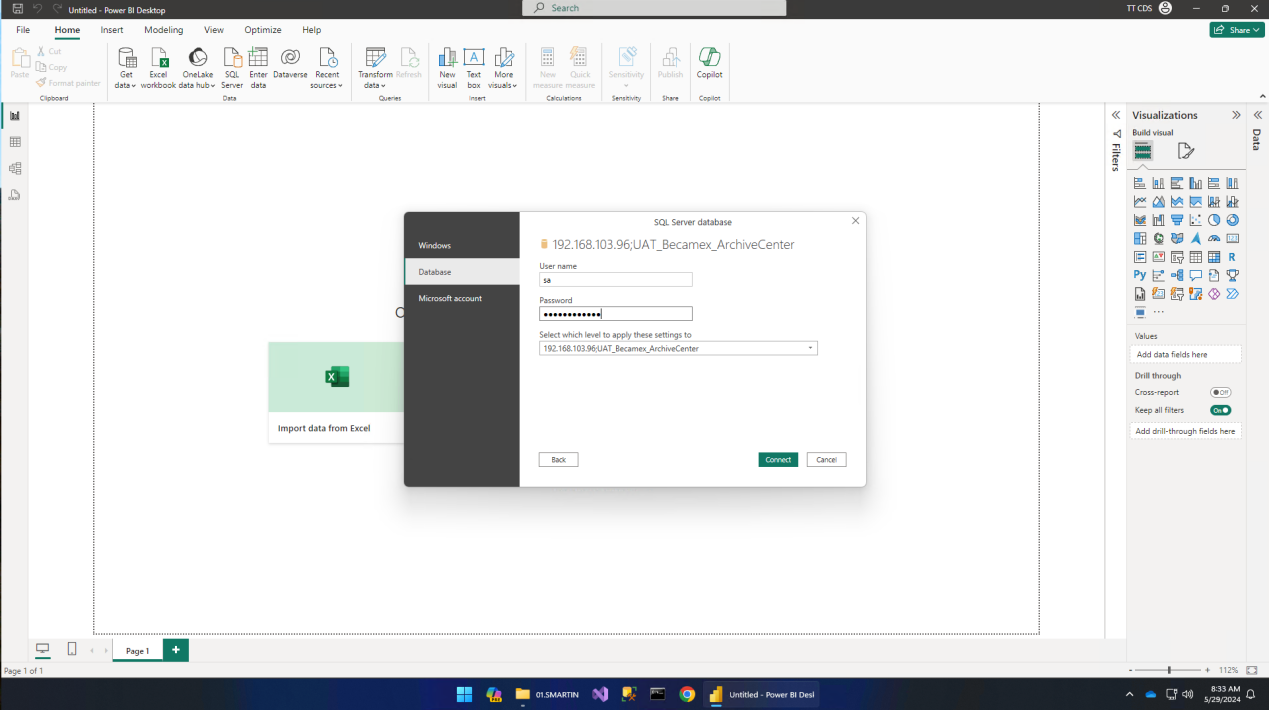
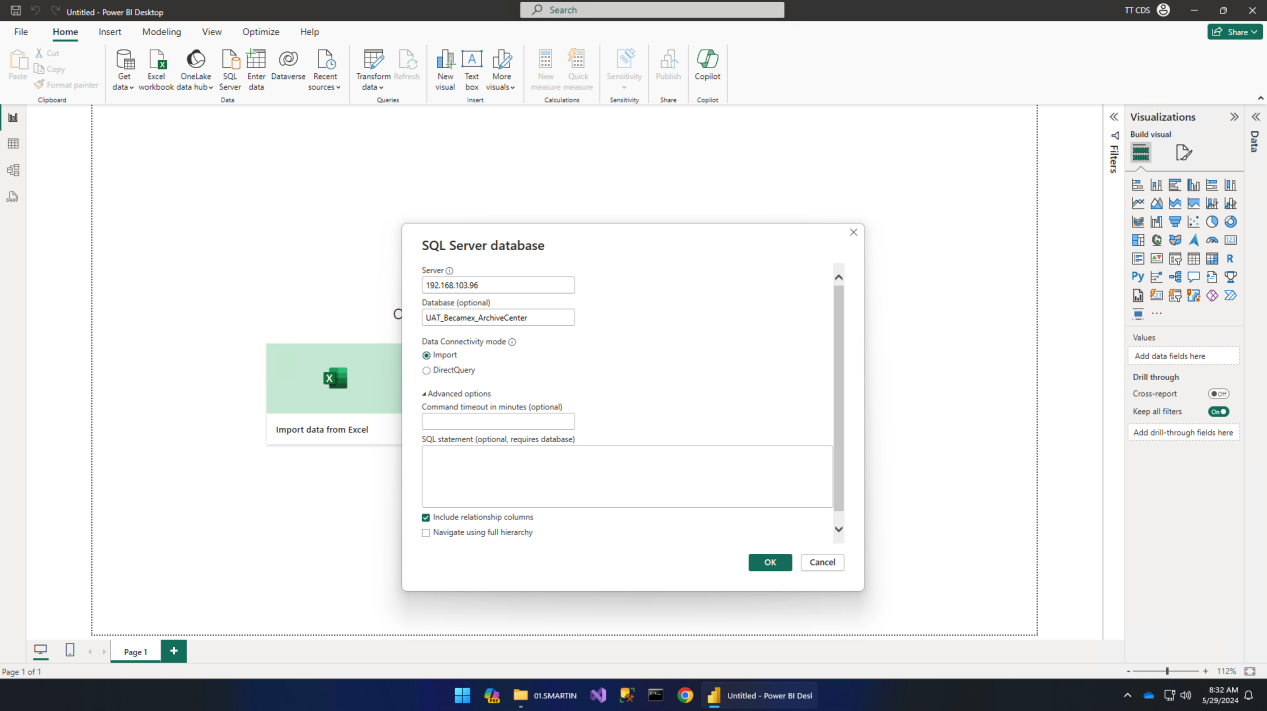
**Bước 4:** Lấy căn bậc hai của MSE để tính RMSE.

**3. Phương pháp đánh giá chung**

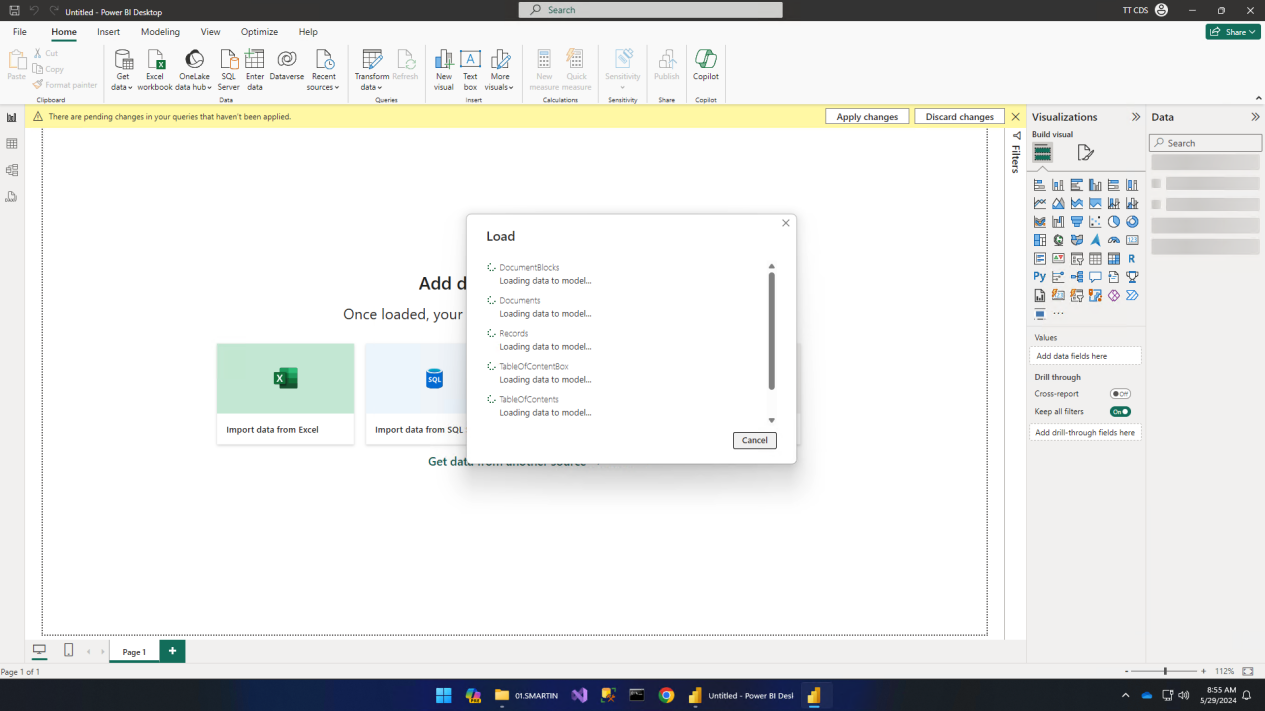
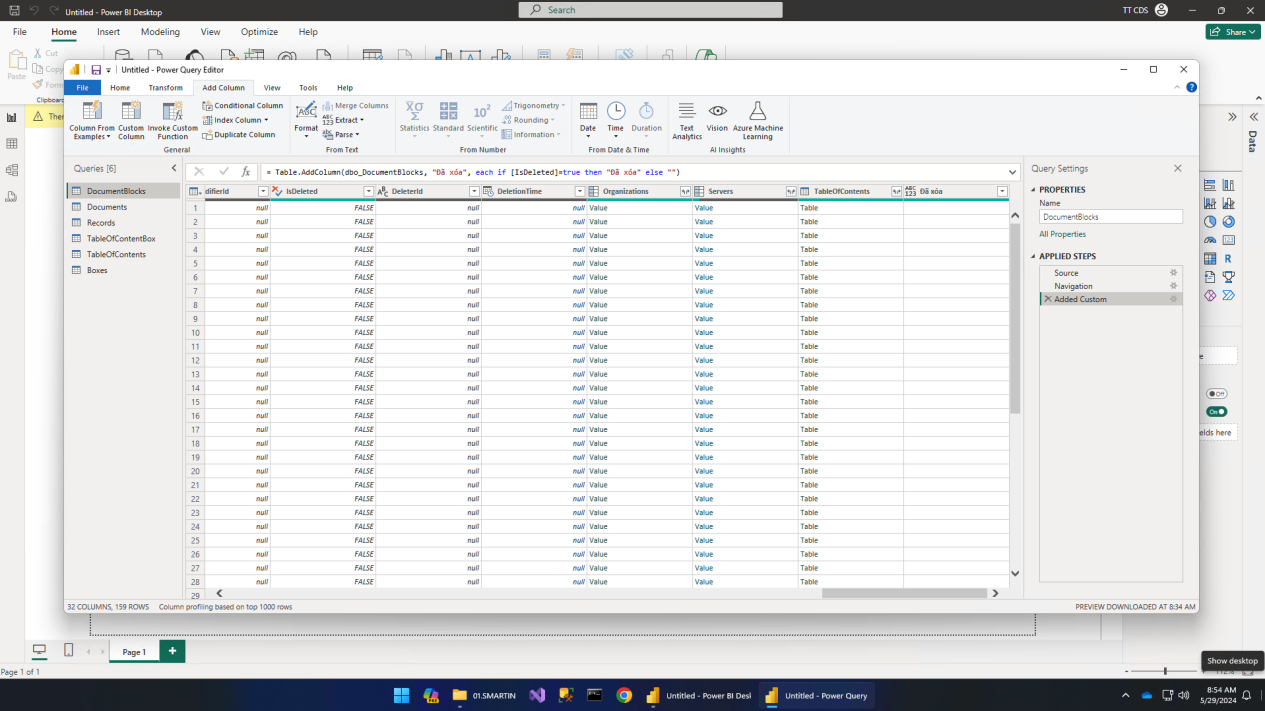
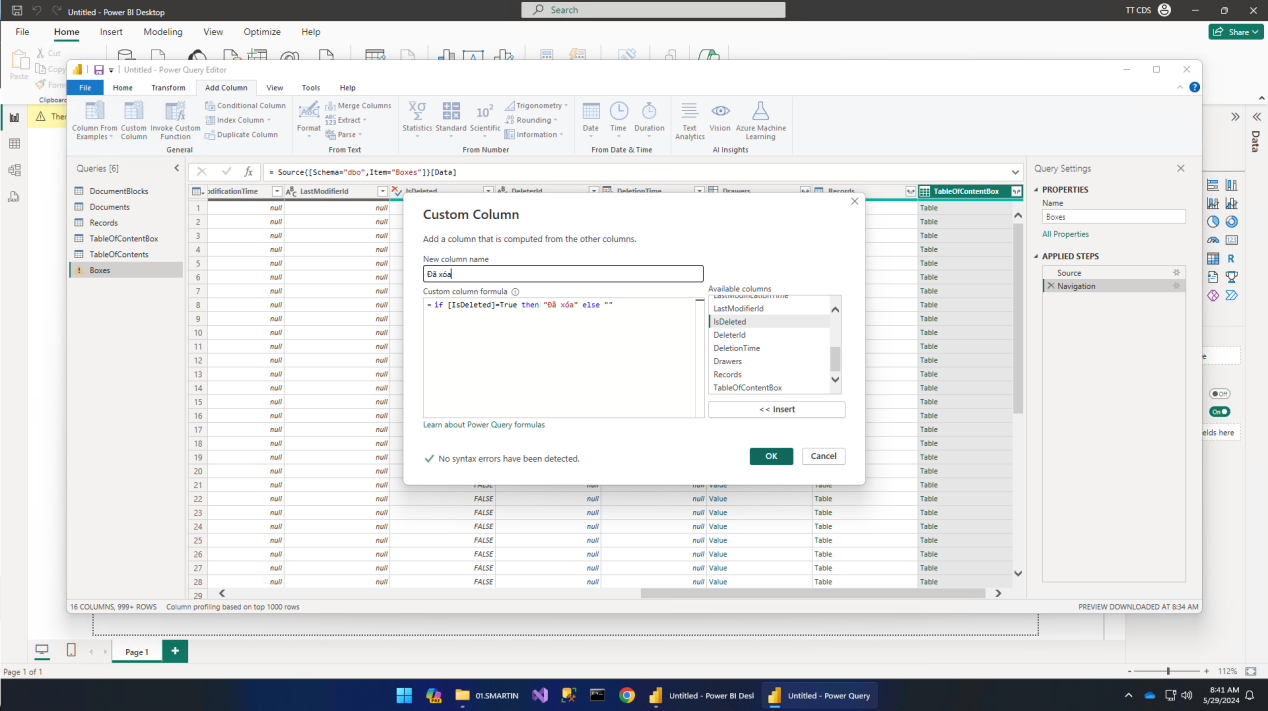
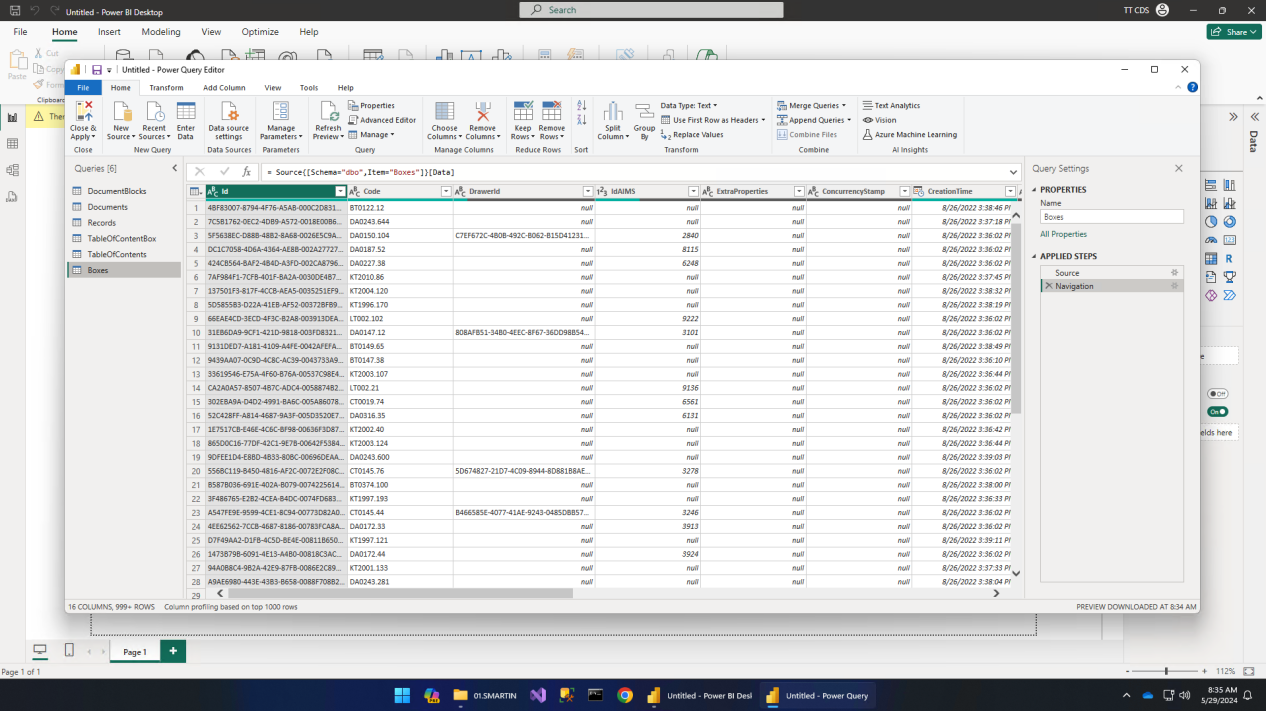
Cross-Validation: Kỹ thuật phân chia dữ liệu thành nhiều tập con và thực hiện huấn luyện và đánh giá trên các tập con đó để đảm bảo mô hình không bị quá khớp (overfitting).

Bootstrap: Kỹ thuật lấy mẫu lại dữ liệu để tạo ra nhiều tập dữ liệu huấn luyện và đánh giá độ chính xác của mô hình trên các tập dữ liệu này.

Holdout Method: Phân chia dữ liệu thành hai phần, một phần dùng để huấn luyện mô hình và phần còn lại dùng để đánh giá mô hình.

Bài tập 2 

Truy vấn dữ liệu



Trực quan hóa dữ liệu

