

## 1. Mô hình hồi quy dự đoán năng suất lúa

a) Các hệ số 0.018, 0.012, 0.004 là hệ số hồi quy biểu thị mức độ thay đổi của biến phụ thuộc Y khi biến độc lập tương ứng tăng thêm 1 đơn vị, với điều kiện các biến còn lại giữ nguyên

b) Ý nghĩa các chỉ số:

- $MSE = 0.045$ : Là trung bình cộng bình phương các sai số (chênh lệch giữa giá trị thực tế và dự đoán). Giá trị càng nhỏ càng tốt.
- $RMSE = 0.212$ : Là căn bậc hai của MSE. Nó đưa đơn vị sai số về cùng đơn vị với biến mục tiêu Y.
- $MAE = 0.185$ : Là trung bình cộng trị tuyệt đối các sai số. Nó cho biết trung bình mô hình dự đoán lệch bao nhiêu so với thực tế.
- $R^2 = 0.941$ : Độ phù hợp của mô hình. Giá trị 0.941 (hay 94.1%) có nghĩa là 94.1% sự biến thiên của năng suất lúa có thể được giải thích bởi 3 biến đầu vào ( $X_1, X_2, X_3$ ).

Trong bài này,  $MSE < 1$  nên lấy căn bậc 2 ra RMSE sẽ luôn được số lớn hơn.

c) Dựa trên các chỉ số:

- $R^2 = 0.941$  gần 1, cho thấy mô hình giải thích dữ liệu cực kỳ tốt.
- $MAE = 0.185$ . So với hệ số chặn là 1.5, mức sai số này là chấp nhận được.

Kết luận: Mô hình này tốt và đáng tin cậy để dự đoán năng suất lúa.

2.

a) Công thức tính hàm Sigmoid:

$$P(Y = 1) = \sigma(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-(5.2 + 0.4X_1 + 0.08X_2 + 0.03X_3)}}$$

b) Precision và Recall trong chẩn đoán bệnh

- Precision: Trong số những người mô hình dự đoán là có bệnh, có bao nhiêu người thực sự có bệnh?
- Recall: Trong số những người thực sự có bệnh, mô hình tìm ra được bao nhiêu người?

Khi nào ưu tiên cái nào?

- Ưu tiên Recall: Khi bỏ sót bệnh nhân gây hậu quả nghiêm trọng (ví dụ: Ung thư, HIV, Covid-19). Thà “chữa” nhầm còn hơn bỏ sót.
- Ưu tiên Precision: Khi việc điều trị rất tốn kém, nguy hiểm hoặc gây ám ảnh tâm lý nặng nề, và bệnh không lây lan nhanh (ví dụ: Phẫu thuật não rủi ro cao, cần chắc chắn mới mổ).

c) Hậu quả của lỗi FP và FN

- FP: Người khỏe nhưng bị báo bệnh. Hậu quả: Bệnh nhân lo lắng, stress, tốn tiền xét nghiệm lại. Ví dụ: Một người bình thường bị chẩn đoán ung thư, họ sẽ cố gắng vay tiền để chữa trị.
- FN: Người bệnh nhưng bị báo khỏe. Hậu quả: Bệnh nhân chủ quan, không điều trị kịp thời, bệnh nặng hơn hoặc tử vong, lây lan cho cộng đồng. Ví dụ: Một người bị ung thư nhưng bác sĩ bảo bình thường -> Bệnh tiến triển xấu và có thể không còn khả năng chữa trị

d) Confusion Matrix:

	Thực tế - Dương tính	Thực tế - Âm tính
Dự đoán - Dương tính	TP = 28	FP = 2
Dự đoán - Âm tính	FN = 82	TN = 88

Tính các chỉ số:

- Accuracy (Độ chính xác tổng thể):

$$\text{Accuracy} = \frac{TP + TN}{\text{Total}} = \frac{28+88}{200} = \frac{116}{200} = 0.58(58\%)$$

Accuracy (58%) không phản ánh đúng bản chất vì dữ liệu hơi mất cân bằng và sự đánh đổi giữa Precision/Recall quá lớn.

- Precision:

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{28}{28+2} = \frac{28}{30} \approx 0.933(93.3\%)$$

Precision rất cao (93.3%): Khi mô hình báo ai đó bị bệnh, khả năng rất cao là họ bị bệnh thật.

- Recall:

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{28}{28+82} = \frac{28}{110} \approx 0.255(25.5\%)$$

Recall quá thấp (25.5%): Mô hình bỏ sót quá nhiều người bệnh (FN = 82 ca). Trong y tế, đây là một mô hình kém vì để lọt lưới phần lớn bệnh nhân.

- F1-score:

$$F1 = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) = 2 * \frac{0.933*0.255}{0.933+0.255} \approx 0.40$$

**Kết luận:** Cần điều chỉnh ngưỡng (threshold) hoặc cải thiện mô hình để tăng Recall lên, chấp nhận giảm Precision xuống một chút.