1. Tiêu đề: Mammographic Mass Data

2. Các thông tin liên quan:

- Chụp nhũ ảnh là phương pháp sàng lọc ung thư vú hiệu quả nhất hiện nay. Tuy nhiên, sự hiệu quả của phương pháp này phụ thuộc không chỉ vào việc đánh giá kết quả chụp quang mà còn phụ thuộc vào kỹ năng và chuyên môn của người thực hiện. Điều này dẫn đến việc thực hiện khoảng 70% ca sinh thiết mà thực tế không cần thiết. Nhằm giảm thiểu số lượng ca sinh thiết vú không cần thiết, trong những năm gần đây, đã được đề xuất một số hệ thống chẩn đoán có sự hỗ trợ của máy tính (CAD). Những hệ thống này giúp các bác sĩ trong quyết định liệu có nên tiến hành sinh thiết vú đối với một dấu hiệu nghi ngờ được phát hiện trong kết quả chụp nhũ ảnh hay thay vào đó tiến hành kiểm tra theo dõi ngắn hạn.

- Tập dữ liệu này có thể được sử dụng để dự đoán mức độ nghiêm trọng (lành tính hoặc ác tính) của tổn thương dựa vào các thuộc tính như đánh giá mức độ BI-RADS (BI-RADS assessment), tuổi bệnh nhân (age) và ba thuộc tính của BI-RADS (hình dạng - shape, bờ - margin, mật độ - density). Trường mức độ nghiêm trọng (severity) đưa ra kết quả cuối cùng là lành tính hoặc ác tính. Tập dữ liệu này được thu thập tại Viện X-quang Đại học Erlangen-Nuremberg từ năm 2003 đến năm 2006, tổng cộng 961 ca, trong đó 516 ca lành tính, 445 ca ác tính. Mỗi trường hợp sẽ có trường đánh giá mức độ BI-RADS (BI-RADS assessment) khác nhau, từ 1 (chắc chắn lành tính) đến 5 (ác tính) được các bác sỹ chỉ định trong quy trình đánh giá kép.

- Tập dữ liệu này được sử dụng để dự đoán mức độ nghiêm trọng (lành tính hoặc ác tính) của khối u vùng vú dựa trên các đặc điểm BI-RADS và tuổi của bệnh nhân. Tập dữ liệu bao gồm: đánh giá BI-RADS, tuổi của bệnh nhân, ba thuộc tính BI-RADS khác (hình dạng, bờ, mật độ) cùng với kết quả thực tế (trường độ nghiêm trọng) cho 516 khối u lành tính và 445 khối u ác tính đã được xác định trên hình ảnh chụp nhũ ảnh số tại Viện X-quang Đại học Erlangen-Nuremberg trong từ năm 2003 đến 2006. Mỗi trường hợp sẽ có trường đánh giá mức độ BI-RADS (BI-RADS assessment) khác nhau, từ 1 (chắc chắn lành tính) đến 5 (ác tính), được các bác sỹ chỉ định trong quy trình đánh giá kép. Giả định rằng tất cả các trường hợp có đánh giá BI-RADS lớn hơn hoặc bằng một giá trị cụ thể (dao động từ 1 đến 5) là ác tính và các trường hợp còn lại là lành tính, ta có thể tính được độ nhạy (sensitivity) và độ đặc hiệu (specificity) tương ứng. Những con số này có thể chỉ ra hiệu suất của hệ thống hỗ trợ chẩn đoán máy tính so với các bác sĩ chuyên khoa.

3. Các thuộc tính trong tập dữ liệu

Tổng cộng 6 thuộc tính (1 trường mục tiêu, 1 thuộc tính không dự đoán, 4 thuộc tính dự đoán)

1. Đánh giá BI-RADS: 1 đến 5 (thứ tự, không dự đoán)

2. Age: tuổi của bệnh nhân tính theo năm (số nguyên)

3. Hình dạng khối (tên danh nghĩa):

1: tròn

2: hình bầu dục

3: thùy

4: không đều

4. Bờ (tên danh nghĩa):

1: đều

2: đa cung nhỏ

3: không rõ

4: không xác định được

5: có gai

5. Mật độ khối (thứ tự):

1: cao

2: iso

3: thấp

4: chứa chất béo

6. Mức độ nghiêm trọng (trường mục tiêu):

0: lành tính

1: ác tính

- BI-RADS là sự đánh giá mức độ tin cậy của việc phân loại mức độ nghiêm trọng và nó không phải là thuộc tính “dự đoán”. Các thuộc tính tuổi, hình dạng, bờ và mật độ là các đặc điểm để xây dựng mô hình và “mức độ nghiêm trọng” là phân loại để dự đoán dựa trên các thuộc tính đó.

Các thuộc tính bị thiếu (missing values):

- Đánh giá BI-RADS: 2

- Tuổi: 5

- Hình dáng: 31

- Bờ: 48

- Mật độ: 76

- Mức độ nghiêm trọng: 0

\* Từ viết tắt và các kiến thức liên quan:

BI-RADS: Breast Imaging Reporting And Data System – là thước đo phân loại theo mức độ tổn thương và nguy cơ ung thư. Bao gồm 6 mức:

- BI-RADS 0: Chưa đầy đủ, cần đánh giá hình ảnh bổ sung và so sánh với các lần chụp nhũ ảnh trước đó (hoặc các xét nghiệm hình ảnh khác).

- BI-RADS 1: Âm tính.

- BI-RADS 2: Sang thương lành tính (0%).

- BI-RADS 3: Sang thương có thể lành tính (≤ 2%), cần theo dõi trong thời gian ngắn.

- BI-RADS 4: Bất thường – đáng nghi ngờ ác tính, tuỳ thuộc vào mức độ sang thương, nên xem xét sinh thiết.

+ BI-RADS 4A: có khả năng mắc ung thư thấp (2% – 10%).

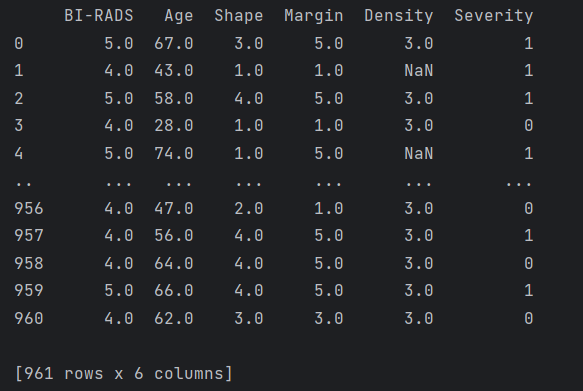
+ BI-RADS 4B: có khả năng mắc ung thư vừa (10% – 50%).

+ BI-RADS 4C: có khả năng mắc ung thư cao (50 – 95%).

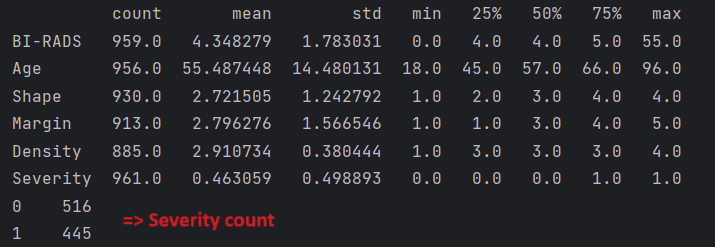
- BI-RADS 5: Khả năng cao là ác tính (≥ 95%), cần thực hiện chẩn đoán sinh thiết.

- Thuật toán sử dụng: KNN, Bayes, Decision Tree, Revisiting SVM (Support Vector Machine).

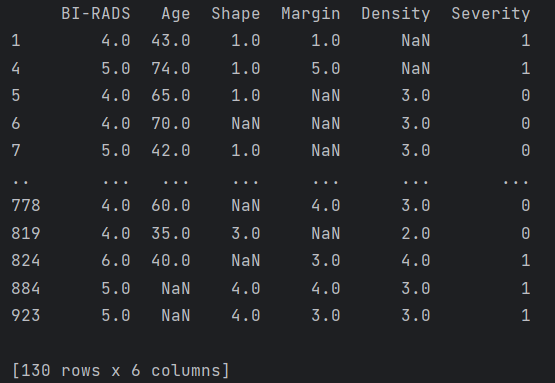
- Tập dữ liệu ban đầu (có chứa missing value): tổng cộng 961 dòng

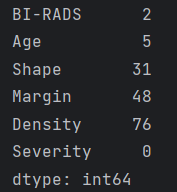


- Hiển thị các thông tin về tập dữ liệu để xác định có nên xử lý dữ liệu trước khi đem đi huấn luyện và kiểm tra:

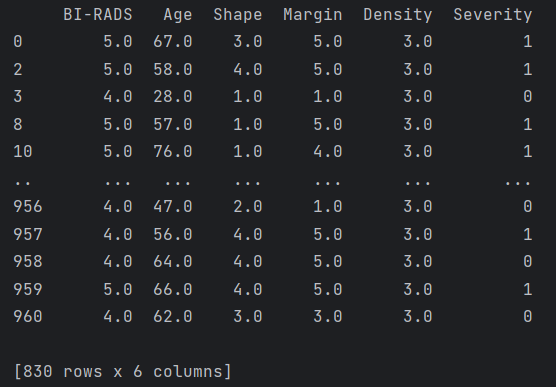


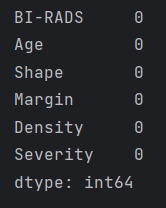
- Hiển thị các dòng dữ liệu chứa missing value: tổng cộng 130 dòng





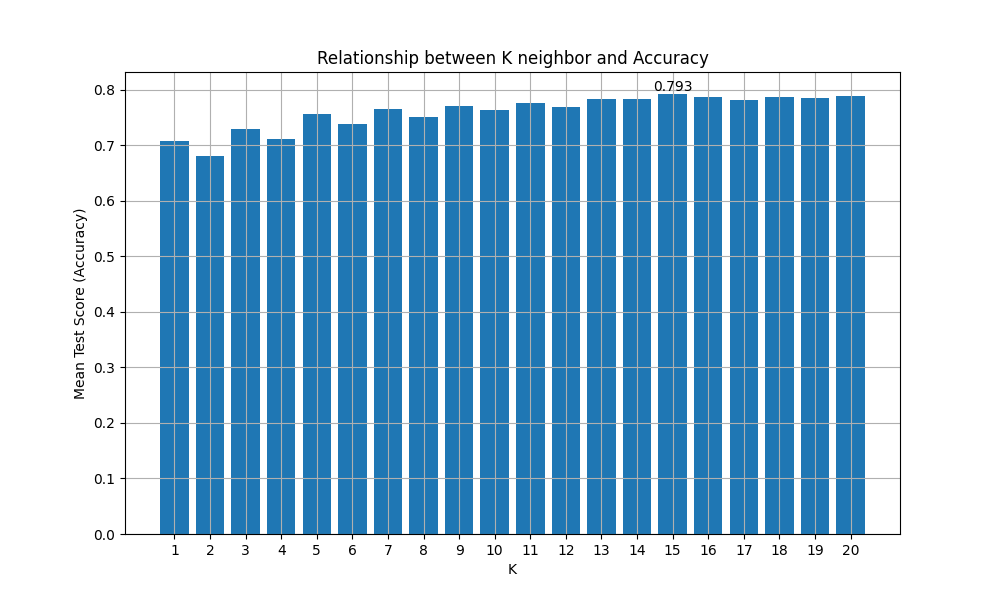
- Tập dữ liệu đã xử lý (bỏ các dòng dữ liệu chứa missing value): còn lại 830 dòng





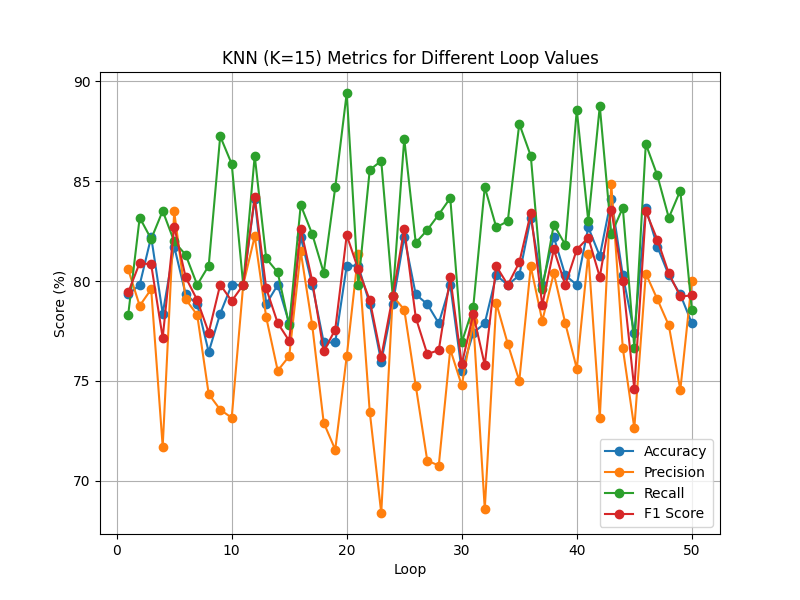
- KNN:

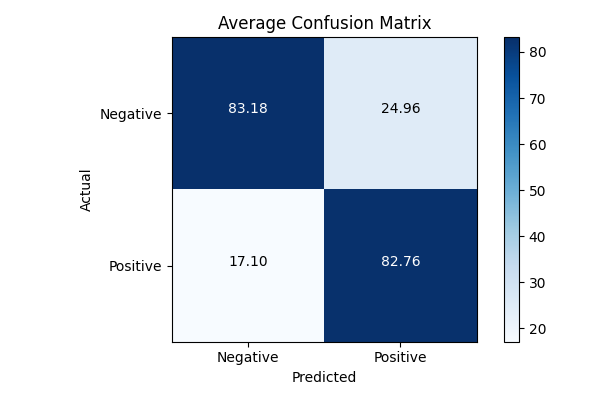
+ Đầu tiên, xác định được K neighbor tốt nhất cho mô hình là K = 15 với độ chính xác là 79.3%.



+ Sử dụng nghi thức hold-out chia tập dữ liệu thành 4 phần, 3 phần được sử dụng để huấn luyện, 1 phần để kiểm tra. Lặp lại quá trình này 50 lần, tính các trung bình các chỉ số để đánh giá mô hình.

+ Thu được chỉ số đánh giá trung bình: Accuracy = 79.779%, Precision = 76.874%, Recall = 82.917%, F1 = 79.692%.

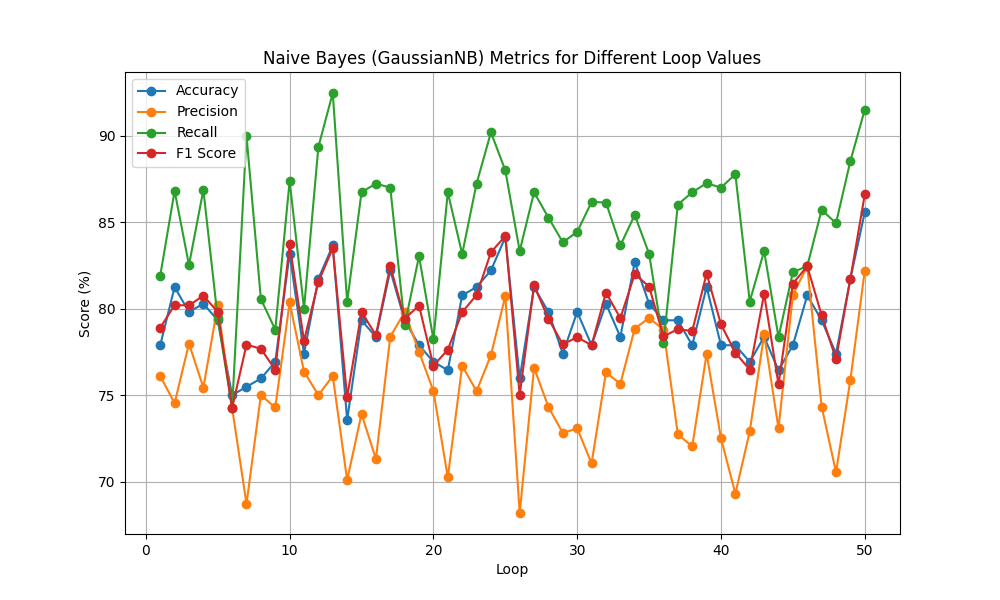


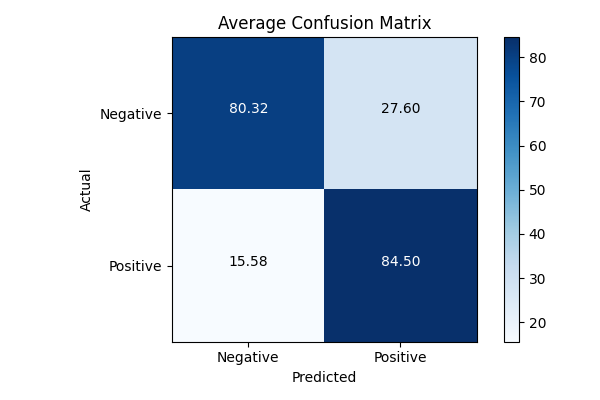


- Bayes:

+ Sử dụng nghi thức hold-out chia tập dữ liệu thành 4 phần, 3 phần được sử dụng để huấn luyện, 1 phần để kiểm tra. Lặp lại quá trình này 50 lần, tính các trung bình các chỉ số để đánh giá mô hình.

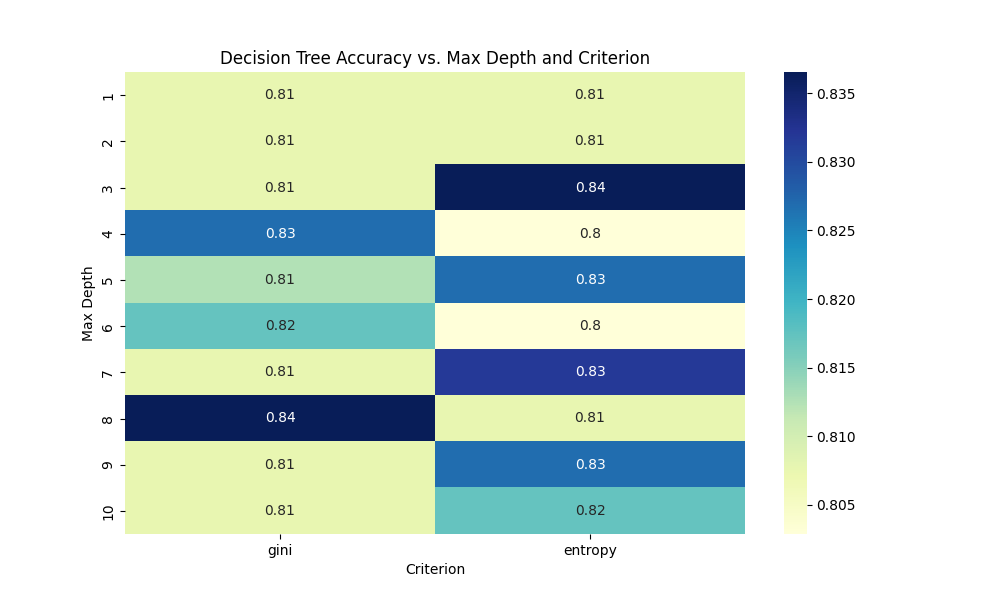
+ Thu được chỉ số đánh giá trung bình: Accuracy = 79.24%, Precision = 75.41%, Recall = 84.514%, F1 = 79.613%.





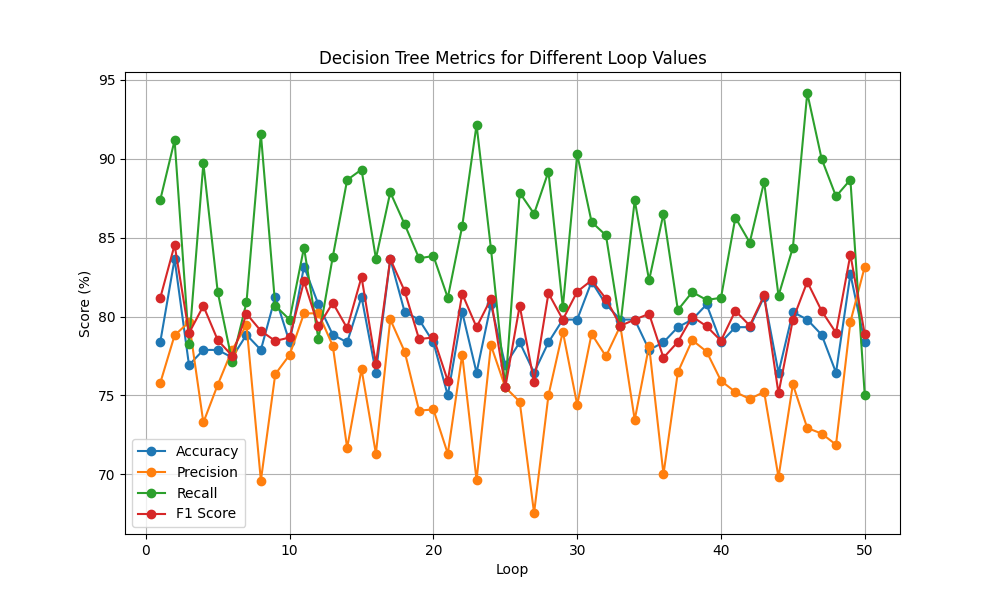
- Decision Tree:

+ Đầu tiên, xác định được max\_depth và criterion tốt nhất cho mô hình là max\_depth = 3 và criterion = entropy với độ chính xác là 84%.

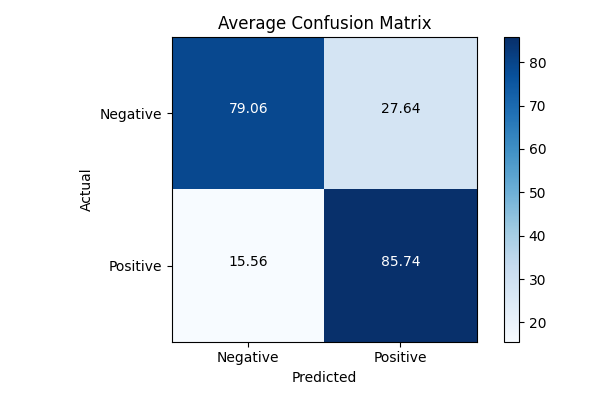


+ Sử dụng nghi thức hold-out chia tập dữ liệu thành 4 phần, 3 phần được sử dụng để huấn luyện, 1 phần để kiểm tra. Sau đó sử dụng DecisionTreeClassifier với max\_depth = 3, criterion = entropy. Lặp lại quá trình này 50 lần, tính các trung bình các chỉ số để đánh giá mô hình.

+ Thu được chỉ số đánh giá trung bình: Accuracy = 79.231%, Precision = 75.756%, Recall = 84.656%, F1 = 79.822%



.

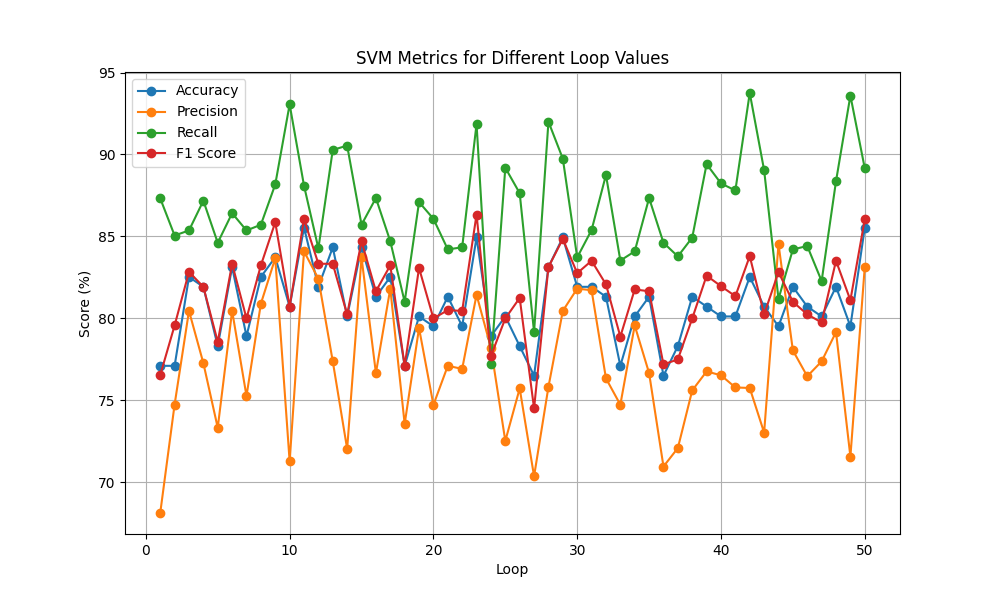


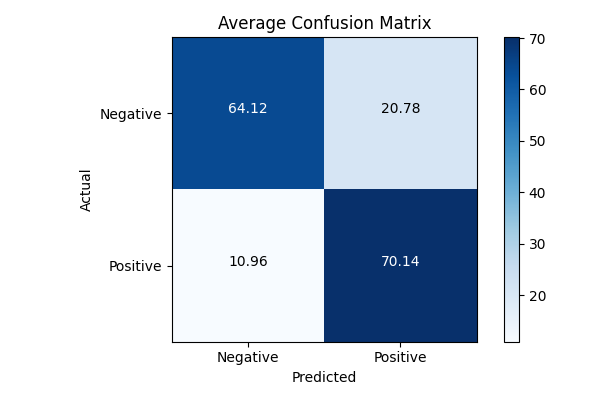
- Revisiting SVM:

+ Đầu tiên, xác định được kernel tốt nhất cho mô hình là linear với độ chính xác là 82.5%.

+ Sử dụng nghi thức hold-out chia tập dữ liệu thành 4 phần, 3 phần được sử dụng để huấn luyện, 1 phần để kiểm tra. Sau đó sử dụng SVC với kernel = linear. Lặp lại quá trình này 50 lần, tính các trung bình các chỉ số để đánh giá mô hình.

+ Thu được chỉ số đánh giá trung bình: Accuracy = 80.88%, Precision = 77.153%, Recall = 86.527%, F1 = 86.527%





- Kết luận: các mô hình đều cho kết quả tương đối tốt, tuy nhiên Revisiting SVM cho kết quả tốt nhất.