SVM

* SVM là một phương pháp phân lớp tuyến tính (linear classifier), với mục đích xác định một siêu phẳng (hyperplane) để phân tách hai lớp của dữ liệu – ví dụ: lớp các ví dụ có nhãn dương (positive) và lớp các ví dụ có nhãn âm (negative)
* Giả sử rằng tập dữ liệu (tập các ví dụ huấn luyện) có thể phân tách được một cách tuyến tính.
* Xét một ví dụ của lớp dương (1) và một ví dụ của lớp âm (,-1) gần nhất đối với siêu phẳng phân tách
* Định nghĩa 2 siêu phẳng lề song song với nhau

* + đi qua và song song với

* + đi qua  và song song với

Sao cho:

* + - nếu
    - nếu

KNN

KNN là thuật toán đi tìm đầu ra của một điểm dữ liệu mới bằng cách *chỉ* dựa trên thông tin của K điểm dữ liệu trong training set gần nó nhất (K-lân cận), *không quan tâm đến việc có một vài điểm dữ liệu trong những điểm gần nhất này là nhiễu*.

* Thường xét k > 1 các ví dụ học (các láng giềng) gần nhất với ví dụ cần phân lớp/dự đoán
* Đối với bài toán phân lớp có 2 lớp, k thường được chọn là một số lẻ, để tránh cân bằng về tỷ lệ các ví dụ giữa 2 lớp

SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) là một phương pháp cân bằng dữ liệu trong các bài toán phân loại mất cân bằng. Phương pháp này tạo thêm các mẫu tổng hợp từ lớp thiểu số (minority class) thay vì chỉ nhân bản ngẫu nhiên, giúp cải thiện hiệu suất của các mô hình học máy khi có sự mất cân bằng giữa các lớp.

**Cách hoạt động của SMOTE**

* **Chọn các điểm dữ liệu từ lớp thiểu số**: SMOTE bắt đầu bằng cách chọn ngẫu nhiên một điểm dữ liệu thuộc lớp thiểu số.
* **Xác định hàng xóm gần nhất**: Sau khi chọn một điểm từ lớp thiểu số, SMOTE tìm k hàng xóm gần nhất của điểm này trong cùng lớp thiểu số. Số lượng hàng xóm k thường được chọn là 5, nhưng có thể điều chỉnh tùy vào tình huống.
* **Tạo các điểm tổng hợp**: Một điểm ngẫu nhiên từ các hàng xóm gần nhất được chọn và SMOTE tạo một điểm tổng hợp mới bằng cách nội suy giữa điểm ban đầu và điểm hàng xóm. Quá trình nội suy được thực hiện bằng cách:

A group of symbols on a white background

Description automatically generated

trong đó  là một số ngẫu nhiên trong khoảng [0,1].

Kết quả là một điểm mới nằm trên đoạn thẳng nối giữa hai điểm, thuộc không gian của lớp thiểu số.

* **Lặp lại**: Quy trình trên được lặp lại cho đến khi đạt được số lượng mẫu mong muốn để cân bằng với lớp đa số.

Minmax normalization

Chuẩn hóa Min-Max (Min-Max Normalization) là một phương pháp được sử dụng trong tiền xử lý dữ liệu nhằm đưa các giá trị của thuộc tính về một khoảng nhất định, thường là từ 0 đến 1.

Giả sử có một tập dữ liệu với một thuộc tính x, và giá trị của thuộc tính này nằm trong khoảng từ ​ đến ​. Chuẩn hóa Min-Max đưa các giá trị của x về khoảng [0,1] bằng công thức:

A white background with black text

Description automatically generated

Maintance

Bộ dữ liệu AI4I (AI for Industry 4.0) là một tập dữ liệu được xây dựng với mục tiêu hỗ trợ việc nghiên cứu và phát triển các giải pháp trí tuệ nhân tạo (AI) trong lĩnh vực công nghiệp, đặc biệt là công nghiệp sản xuất thông minh. Bộ dữ liệu này chứa các thông tin liên quan đến các quá trình sản xuất và hoạt động của máy móc trong môi trường sản xuất công nghiệp, bao gồm các chỉ số về trạng thái hoạt động, hiệu suất, và các yếu tố gây lỗi của thiết bị.

Bộ dữ liệu AI4I có các cột thông tin như sau:

* **UDI** (Unique Device Identifier): Mã định danh duy nhất của thiết bị, giúp nhận diện từng mẫu dữ liệu một cách riêng biệt.
* **Product ID**: Mã sản phẩm được sản xuất, có thể liên quan đến loại sản phẩm hoặc lô sản phẩm cụ thể.
* **Type**: Loại sản phẩm hoặc danh mục của sản phẩm, phân biệt các dòng sản phẩm khác nhau trong quá trình sản xuất.
* **Air Temperature [K]**: Nhiệt độ không khí xung quanh máy móc trong quá trình sản xuất, tính bằng độ Kelvin (K).
* **Process Temperature [K]**: Nhiệt độ tại điểm hoạt động của quá trình sản xuất, tính bằng độ Kelvin (K). Thường cao hơn nhiệt độ không khí do phát sinh nhiệt từ hoạt động máy móc.
* **Rotational Speed [rpm]**: Tốc độ quay của động cơ, tính bằng vòng/phút (rpm), thể hiện tốc độ hoạt động của máy móc.
* **Torque [Nm]**: Mô-men xoắn (torque), tính bằng Newton-mét (Nm), thể hiện lực quay của máy móc trong quá trình vận hành.
* **Tool Wear [min]**: Độ mòn của dụng cụ sản xuất, tính bằng phút (min), biểu thị thời gian công cụ đã được sử dụng trong sản xuất. Đây là thông số quan trọng để đánh giá tình trạng của công cụ và khả năng cần bảo trì hoặc thay thế.
* **Target**: Nhãn mục tiêu (target), xác định xem một sản phẩm có đạt tiêu chuẩn hay không. Đây có thể là biến nhị phân, cho biết sản phẩm đã qua kiểm tra chất lượng (đạt hoặc không đạt).
* **Failure Type**: Loại lỗi gặp phải (nếu có) trong quá trình sản xuất. Các giá trị trong cột này có thể bao gồm:
  + Không lỗi (No Failure)
  + Lỗi quá tải công cụ (Tool Wear Failure)
  + Lỗi nguồn nhiệt (Heat Dissipation Failure)
  + Lỗi công cụ (Power Failure)
  + Lỗi bôi trơn (Overstrain Failure)
  + Lỗi môi trường (Random Failures)

Bộ dữ liệu này có thể được dùng để xây dựng các mô hình AI và phân tích dữ liệu, nhằm dự đoán lỗi hoặc tối ưu hóa hiệu suất của máy móc. Các cột **Rotational Speed**, **Torque**, và **Tool Wear** có thể đóng vai trò quan trọng trong việc phân tích sự hao mòn của máy móc và dự đoán thời điểm bảo trì cần thiết. Các cột **Air Temperature** và **Process Temperature** giúp đánh giá sự ảnh hưởng của nhiệt độ đến hiệu suất hoạt động và tuổi thọ của thiết bị.

SVC

1. **Ma trận nhầm lẫn**:
   * Mô hình SVC đã phân loại chính xác phần lớn các mẫu trong tập dữ liệu, với 2517 và 2502 là số lượng dự đoán đúng cho mỗi lớp.
   * Tuy nhiên, vẫn còn có một số dự đoán sai, với 364 mẫu của lớp 0 bị dự đoán thành lớp 1 và 316 mẫu của lớp 1 bị dự đoán thành lớp 0.
2. **Biểu đồ ROC – AUC**:
   * Đường cong ROC gần sát với góc trên bên trái của biểu đồ, cho thấy mô hình có khả năng phân biệt tốt giữa các lớp.
   * Giá trị AUC đạt 0.95, là một con số cao, chứng tỏ mô hình có độ nhạy cao trong việc phân biệt các lớp.
3. **Biểu đồ Precision-Recall**:
   * Đường cong Precision-Recall cho thấy mô hình đạt độ chính xác và độ nhạy cao khi tỷ lệ Recall ở mức cao.
   * Diện tích dưới đường cong (AP = 0.94) cũng cho thấy khả năng của mô hình trong việc xử lý lớp thiểu số hoặc mất cân bằng trong dữ liệu.
4. **Hiệu suất của mô hình**:
   * Các chỉ số đánh giá của mô hình như Precision (0.869), Recall (0.881), F1-score (0.875), và Accuracy (0.874) đều cho thấy mô hình có hiệu suất tốt và ổn định.
   * Precision và Recall có giá trị khá gần nhau, điều này cho thấy mô hình có sự cân bằng giữa việc nhận diện đúng các mẫu và giảm thiểu dự đoán sai.

**Tổng kết:**

Mô hình SVC này đạt hiệu suất khá tốt trên tập dữ liệu hiện tại. Tuy nhiên, vẫn có thể tối ưu thêm để giảm số lượng mẫu bị phân loại sai, chẳng hạn bằng cách điều chỉnh các tham số của mô hình hoặc sử dụng các kỹ thuật tăng cường dữ liệu nếu có sự mất cân bằng giữa các lớp.

KNN

1. **Ma trận nhầm lẫn**:
   * Mô hình KNN phân loại đúng phần lớn các mẫu trong tập dữ liệu, với 2727 và 2857 là số lượng dự đoán đúng cho mỗi lớp.
   * Tuy nhiên, vẫn có một số lượng nhỏ mẫu bị phân loại sai, với 174 mẫu của lớp 0 bị dự đoán thành lớp 1 và 39 mẫu của lớp 1 bị dự đoán thành lớp 0, cho thấy mô hình có hiệu suất tốt hơn ở lớp 1.
2. **Biểu đồ ROC – AUC**:
   * Đường cong ROC sát với góc trên bên trái của biểu đồ, cho thấy mô hình có khả năng phân biệt giữa các lớp khá tốt.
   * Giá trị AUC đạt 0.99, cho thấy mô hình KNN có độ chính xác cao trong việc phân loại và khả năng phân biệt hai lớp mạnh mẽ.
3. **Biểu đồ Precision-Recall**:
   * Đường cong Precision-Recall cho thấy mô hình duy trì độ chính xác và độ nhạy cao trong phần lớn các giá trị Recall.
   * Diện tích dưới đường cong Precision-Recall (AP = 0.98) cho thấy mô hình hoạt động hiệu quả trong việc phân loại các mẫu của lớp thiểu số.
4. **Hiệu suất của mô hình**:
   * Các chỉ số Precision (0.943), Recall (0.987), F1-score (0.964), và Accuracy (0.963) đều cho thấy mô hình KNN đạt hiệu suất rất cao và ổn định.
   * Giá trị Recall cao (0.987) cho thấy mô hình ít bỏ sót các mẫu dương tính, trong khi Precision cũng cao, cho thấy mô hình ít dự đoán sai các mẫu âm tính.

**Tổng kết:**

Mô hình KNN cho thấy hiệu suất rất tốt trên tập dữ liệu hiện tại, với độ chính xác và khả năng phân loại rất cao. Các chỉ số đều ở mức tốt, đặc biệt là Recall và F1-score, cho thấy mô hình có thể cân bằng giữa độ chính xác và độ nhạy.

So sánh

| **Chỉ số** | **Mô hình SVC** | **Mô hình KNN** | **Nhận xét** |
| --- | --- | --- | --- |
| **Precision** | 0.869 | 0.943 | KNN có độ chính xác cao hơn SVC, ít dự đoán sai hơn cho các mẫu dương tính. |
| **Recall** | 0.881 | 0.987 | KNN có Recall cao hơn, cho thấy khả năng nhận diện đúng các mẫu dương tính tốt hơn. |
| **F1-score** | 0.875 | 0.964 | KNN đạt F1-score cao hơn, cho thấy cân bằng tốt giữa Precision và Recall. |
| **Accuracy** | 0.874 | 0.963 | Độ chính xác của KNN cao hơn SVC, cho thấy hiệu suất tổng thể tốt hơn trên tập dữ liệu. |
| **MAE** | 0.1256 | 0.0367 | MAE của KNN thấp hơn, nghĩa là sai số trung bình của KNN nhỏ hơn, dự đoán chính xác hơn. |
| **MSE** | 0.1256 | 0.0367 | MSE của KNN cũng thấp hơn, cho thấy sai số bình phương nhỏ hơn so với SVC. |
| **R-squared (R²)** | 0.4977 | 0.8530 | KNN có R² cao hơn nhiều, nghĩa là giải thích tốt hơn phần lớn biến thiên của dữ liệu. |

**Tổng kết:**

* Mô hình **KNN** thể hiện hiệu suất vượt trội hơn so với **SVC** trên tất cả các chỉ số đánh giá, đặc biệt là ở Precision, Recall, F1-score, và R².
* Với độ chính xác cao hơn, sai số thấp hơn, và khả năng giải thích biến thiên dữ liệu tốt hơn, **KNN** là lựa chọn tối ưu cho bài toán

Material

Bộ dữ liệu gồm 2 bảng: material.csv và Data.csv, gồm 15 đặc trưng mô tả các đặc điểm kĩ thuật của vật liệu và 1 nhãn.

Dưới đây là tổng quan về các đặc trưng:

* **Std**: Tiêu chuẩn vật liệu hoặc mã định danh tiêu chuẩn (Standard).
* **ID**: Mã định danh duy nhất của vật liệu.
* **Material**: Loại vật liệu.
* **Heat treatment**: Phương pháp xử lý nhiệt của vật liệu.
* **Su**: Độ bền kéo (Ultimate tensile strength).
* **Sy**: Giới hạn chảy (Yield strength).
* **A5**: Độ giãn dài đến khi đứt (percentage elongation).
* **Bhn**: Độ cứng Brinell của vật liệu.
* **E**: Mô đun đàn hồi (Elastic modulus).
* **G**: Mô đun cắt (Shear modulus).
* **mu**: Hệ số ma sát (Friction coefficient).
* **Ro**: Khối lượng riêng (Density).
* **pH**: Độ pH, có thể liên quan đến tính chất hóa học bề mặt hoặc môi trường thử nghiệm.
* **Desc**: Mô tả vật liệu hoặc chi tiết về mẫu thử.
* **HV**: Độ cứng Vickers.
* **Use**: Vật liệu có được sử dụng hay không

SVC

* **Ma trận nhầm lẫn**:
  + Mô hình có tổng cộng 319 dự đoán chính xác cho nhãn 0 và 378 dự đoán chính xác cho nhãn 1.
  + Có 78 trường hợp nhãn 0 được dự đoán sai thành nhãn 1.
  + Không có trường hợp nào mà nhãn 1 bị dự đoán sai thành nhãn 0, cho thấy recall cho nhãn 1 là hoàn hảo.
* **Biểu đồ ROC-AUC**:
  + Đường ROC có giá trị AUC là 0.933, cho thấy mô hình phân biệt tốt giữa hai lớp.
  + Đường cong gần như sát trục y, chứng tỏ mô hình có khả năng nhận diện tốt các mẫu dương tính thật so với mẫu âm tính.
* **Biểu đồ Precision-Recall**:
  + Đường Precision-Recall cho thấy mô hình duy trì độ chính xác cao ngay cả khi recall tăng lên.
  + Với AP (Average Precision) là 0.82, biểu đồ này cho thấy mô hình hoạt động tốt cho cả hai chỉ số precision và recall.
* **Hiệu suất tổng thể của mô hình**:
  + Precision: 0.829, cho thấy khoảng 82.9% các dự đoán dương tính là chính xác.
  + Recall: 1.000, tức là mô hình phát hiện hoàn toàn các trường hợp dương tính thật.
  + F1-score: 0.906, là một chỉ số cân bằng giữa precision và recall.
  + Accuracy: 0.899, cho thấy độ chính xác tổng thể của mô hình đạt 89.9%.

**Nhận xét tổng quát**

Mô hình SVC này có độ chính xác và recall rất cao, đặc biệt là không bỏ sót bất kỳ trường hợp dương tính nào, thể hiện qua recall đạt 1.000. Tuy nhiên, vẫn có một số dự đoán sai đối với nhãn 0, khiến cho precision thấp hơn. Điều này có thể được cải thiện thêm nếu tìm cách tối ưu hóa precision mà không làm giảm recall.

KNN

1. **Ma trận nhầm lẫn**:
   * Mô hình dự đoán đúng 370 trường hợp nhãn 0 và 368 trường hợp nhãn 1.
   * Có 27 trường hợp nhãn 0 được dự đoán sai thành nhãn 1, và 10 trường hợp nhãn 1 bị dự đoán sai thành nhãn 0.
   * Điều này cho thấy mô hình có khả năng phân biệt hai lớp khá tốt, nhưng vẫn có một số dự đoán sai ở cả hai nhãn.
2. **Biểu đồ ROC-AUC**:
   * Đường ROC có giá trị AUC là 0.98, cho thấy khả năng phân biệt tốt giữa hai lớp.
   * Đường cong ROC gần sát trục y và đạt giá trị cao, minh chứng rằng mô hình có hiệu quả cao trong việc nhận diện các mẫu dương tính và âm tính.
3. **Biểu đồ Precision-Recall**:
   * Biểu đồ này cho thấy mô hình duy trì precision cao ngay cả khi recall tăng lên, và đạt AP (Average Precision) là 0.97.
   * Đường Precision-Recall tương đối cao và ổn định, cho thấy mô hình duy trì sự cân bằng giữa precision và recall tốt.
4. **Hiệu suất tổng thể của mô hình**:
   * Precision: 0.932, cho thấy 93.2% các dự đoán dương tính là chính xác.
   * Recall: 0.974, nghĩa là mô hình phát hiện được 97.4% các trường hợp dương tính.
   * F1-score: 0.952, là một chỉ số cân bằng giữa precision và recall.
   * Accuracy: 0.952, cho thấy độ chính xác tổng thể của mô hình đạt 95.2%.

**Nhận xét tổng quát**

Mô hình KNN có hiệu suất cao với độ chính xác, precision, recall, và F1-score đều đạt giá trị cao. Giá trị AUC cũng cao, cho thấy mô hình phân biệt tốt giữa hai lớp. Tuy nhiên, mô hình vẫn có một số nhầm lẫn, đặc biệt là đối với nhãn 0 khi bị nhầm thành nhãn 1. Điều này có thể được cải thiện thêm bằng cách điều chỉnh các tham số của KNN hoặc thử các kỹ thuật giảm lỗi dương tính và âm tính giả.

So sánh

**Nhận xét chi tiết**

1. **Precision**: KNN có precision cao hơn (0.932 so với 0.829 của SVC). Điều này có nghĩa là mô hình KNN ít dự đoán nhầm các trường hợp dương tính giả hơn SVC.
2. **Recall**: SVC có recall cao hơn, đạt giá trị hoàn hảo 1.0, trong khi KNN đạt 0.974. Điều này cho thấy SVC phát hiện tất cả các trường hợp dương tính, nhưng KNN bỏ sót một số ít trường hợp dương tính.
3. **F1-score**: KNN có F1-score cao hơn (0.952 so với 0.906), chứng tỏ KNN duy trì sự cân bằng tốt hơn giữa precision và recall.
4. **Accuracy**: KNN có độ chính xác tổng thể cao hơn (0.952 so với 0.899 của SVC), cho thấy mô hình KNN thực hiện dự đoán chính xác hơn trên toàn bộ tập dữ liệu.
5. **AUC (ROC):** KNN có AUC cao hơn (0.980 so với 0.933 của SVC), điều này chỉ ra rằng KNN có khả năng phân biệt tốt hơn giữa hai lớp.
6. **AP (Precision-Recall):** KNN có AP cao hơn (0.97 so với 0.82 của SVC), cho thấy KNN duy trì độ chính xác cao ngay cả khi recall tăng lên.

**Kết luận**

Mô hình KNN có hiệu suất tổng thể tốt hơn so với SVC ở hầu hết các chỉ số, đặc biệt là về độ chính xác, precision, F1-score, và AUC. Tuy nhiên, SVC đạt recall cao hơn, nghĩa là nó phát hiện tất cả các mẫu dương tính mà không bỏ sót, điều này có thể phù hợp hơn trong các ứng dụng nhạy cảm khi cần hạn chế âm tính giả.