MỤC LỤC

[**1.** **Giới thiệu** 2](#_Toc183389886)

[**2.** **Bộ dữ liệu material** 2](#_Toc183389887)

[**3.** **Tiền xử lý dữ liệu** 2](#_Toc183389888)

[**4.** **Cân bằng dữ liệu SMOTE - Synthetic Minority Over-sampling Technique** 3](#_Toc183389889)

[**5.** **Chuẩn hóa Minmax** 4](#_Toc183389890)

[**6.** **Mối tương quan giữa các đặc trưng** 4](#_Toc183389891)

[**7.** **Triển khai trên tập dữ liệu** 6](#_Toc183389892)

[**8.** **Phân tích hiệu suất của các thuật toán** 8](#_Toc183389893)

[**a.** **Support Vector Classification - SVC** 9](#_Toc183389894)

[**b.** **K-nearest Neighbor - KNN** 10](#_Toc183389895)

[**9.** **Thảo luận** 11](#_Toc183389896)

1. **Giới thiệu**

Lựa chọn vật liệu trong ngành ô tô là một yếu tố quan trọng, ảnh hưởng trực tiếp đến hiệu suất, độ an toàn, chi phí và tác động môi trường của phương tiện. Các nhà sản xuất ô tô luôn tìm cách sử dụng vật liệu tối ưu để đáp ứng yêu cầu về độ bền, trọng lượng nhẹ, khả năng chống ăn mòn và chi phí thấp. Với sự phát triển công nghệ, ngành ô tô đang dần chuyển sang các vật liệu nhẹ, bền và thân thiện môi trường, nhằm tạo ra những mẫu xe không chỉ an toàn và tiết kiệm nhiên liệu mà còn có khả năng tái chế cao, giảm thiểu tác động đến môi trường.

Ứng dụng học máy (Machine Learning - ML) vào việc lựa chọn vật liệu trong ngành ô tô là một bước tiến lớn giúp tối ưu hóa quy trình, giảm thời gian phát triển sản phẩm và nâng cao hiệu suất. ML giúp phân tích và dự đoán các đặc tính của vật liệu, từ đó đưa ra lựa chọn phù hợp với yêu cầu kỹ thuật, chi phí và tính bền vững. Bằng cách tận dụng ML để dự đoán tính chất, tối ưu hóa chi phí và tìm ra vật liệu mới thân thiện môi trường, ngành ô tô có thể cải thiện chất lượng, độ an toàn và tính bền vững của sản phẩm.

Trong tương lai, các công nghệ ML kết hợp với Digital Twin và các phương pháp mô phỏng sẽ mở ra nhiều cơ hội lớn trong việc phát triển các loại vật liệu tiên tiến cho ô tô.

1. **Bộ dữ liệu 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

1. **Tiền xử lý dữ liệu**

Tiền xử lý dữ liệu là một bước rất quan trọng trong việc giải quyết bất kỳ vấn đề nào trong lĩnh vực Học Máy. Hầu hết các bộ dữ liệu được sử dụng trong các vấn đề liên quan đến Học Máy cần được xử lý, làm sạch và biến đổi trước khi một thuật toán Học Máy có thể được huấn luyện trên những bộ dữ liệu này.

Việc xử lý dữ liệu trên tập dữ liệu phần mềm độc hại TUANDROMD tập trung vào việc rời rạc hóa, các giá trị bị thiếu được quy kết và dữ liệu không cân bằng, loại bỏ các biến chất lượng thấp và sử dụng các quy trình làm sạch dữ liệu để xác định và sửa chữa sự không nhất quán trong tập dữ liệu. Việc xác định và loại bỏ các mẫu trùng lặp khỏi tập dữ liệu là một bước quan trọng trong giai đoạn chuẩn bị dữ liệu, bằng cách sử dụng phương pháp sau. Giả sử P và H lần lượt là tên gói và giá trị băm SHA256 của mỗi mẫu. Tập hợp các mẫu duy nhất có thể thu được bằng cách chọn mẫu s thuộc S sao cho s có sự kết hợp duy nhất của P và H, như minh họa dưới đây:



Xử lý các giá trị đặc trưng bị thiếu bằng cách sử dụng phép quy ước trung bình hoặc trung vị dựa trên phân phối của các giá trị đặc trưng.

A close up of a number

Description automatically generated with medium confidence

Hoặc

A close up of a text

Description automatically generated

1. **Cân bằng dữ liệu SMOTE - Synthetic Minority Over-sampling Technique**

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.
* **Borderline-SMOTE**

Borderline-SMOTE (Borderline Synthetic Minority Over-sampling Technique) là một biến thể của thuật toán SMOTE (Synthetic Minority Over-sampling Technique), được sử dụng để cân bằng dữ liệu mất cân bằng trong các bài toán phân loại. Trong các bài toán này, số lượng mẫu của lớp thiểu số (minority class) thường nhỏ hơn rất nhiều so với lớp đa số (majority class), làm giảm độ chính xác và độ tin cậy của mô hình. Borderline-SMOTE tập trung vào việc tạo mẫu tổng hợp cho các điểm dữ liệu nằm gần biên giới giữa các lớp, nơi mà khả năng nhầm lẫn giữa các lớp thường cao hơn.

**Cách hoạt động của Borderline-SMOTE**

1. **Xác định các mẫu khó phân loại**: Đầu tiên, Borderline-SMOTE xác định các mẫu của lớp thiểu số nằm gần biên giới phân chia giữa lớp thiểu số và lớp đa số, cụ thể là các mẫu có nguy cơ bị nhầm lẫn cao hơn. Các mẫu này thường nằm gần các điểm của lớp đa số.
2. **Phân loại các điểm biên**: Những điểm dữ liệu thuộc lớp thiểu số gần kề với lớp đa số được chọn làm các điểm biên giới (borderline). Borderline-SMOTE xác định những điểm này bằng cách sử dụng một số tiêu chí như số lượng hàng xóm gần nhất thuộc lớp đa số. Những mẫu này có nguy cơ cao bị phân loại sai trong quá trình học của mô hình.
3. **Tạo các mẫu tổng hợp**: Sau khi xác định các điểm biên giới, Borderline-SMOTE tạo ra các điểm tổng hợp mới bằng cách nội suy giữa các điểm biên này với các hàng xóm gần nhất của chúng (thuộc lớp thiểu số). Điều này giúp tăng số lượng mẫu ở các khu vực quan trọng nhất (biên giới giữa các lớp) thay vì tăng số lượng mẫu trên toàn bộ không gian lớp thiểu số như SMOTE gốc.
4. **Chuẩn hóa Minmax**

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

Chuẩn hóa Minimax đặc biệt hữu ích trong ML, nơi các đặc điểm có quy mô hoặc phạm vi lớn có thể chi phối quá trình phân tích và tạo ra kết quả không chính xác. Bằng cách áp dụng chuẩn hóa Minimax cho từng đặc điểm trong tập dữ liệu, chúng ta có thể đảm bảo rằng tất cả các đặc điểm đều có đóng góp như nhau cho quá trình phân tích, bất kể quy mô hoặc phạm vi ban đầu của chúng.

1. **Mối tương quan giữa các đặc trưng**

Ma trận tương quan là một công cụ thống kê được sử dụng để đo lường và thể hiện mức độ tương quan giữa các biến trong một tập dữ liệu.

Trong ma trận này:

* Màu đỏ đậm đại diện cho mối tương quan dương cao (gần +1), có nghĩa là khi một biến tăng, biến kia cũng có xu hướng tăng.
* Màu xanh đậm đại diện cho mối tương quan âm cao (gần -1), có nghĩa là khi một biến tăng, biến kia có xu hướng giảm.
* Màu xám hoặc màu nhạt gần như không có mối tương quan (giá trị gần 0).

Dưới đây là ma trận tương quan thể hiện mối tương quan giữa các đặc trưng của bộ dữ liệu:

A screenshot of a graph

Description automatically generated

* Các cặp như (Su, Sy) hoặc (A5, E) có vẻ có độ tương quan cao, thể hiện qua màu sắc đậm hơn. Điều này có thể cho thấy rằng khi một biến tăng, biến kia có xu hướng tăng (hoặc giảm) tương ứng.
* Các biến như category và material\_type có tương quan thấp với hầu hết các đặc tính vật liệu khác, điều này có thể ngụ ý rằng các thuộc tính vật lý và cơ học của vật liệu không bị ảnh hưởng nhiều bởi danh mục hoặc loại vật liệu được chỉ định.

1. **Triển khai trên tập dữ liệu**

**A black and yellow text

Description automatically generated**

* Loại bỏ các đặc trưng tên gọi

A close up of text

Description automatically generated

A screenshot of a computer code

Description automatically generated

* Phân loại đặc trưng “Material” từng mẫu thành các loại vật liệu như
  + Thép
  + Đồng
  + Bạc
  + Sắt
  + Nhôm
  + Magie

A screen shot of a computer program

Description automatically generated

* Phân loại đặc trưng Heat treatment thành các loại phương pháp xử lý nhiệt như:
  + Gia Công Nhiệt
  + Gia Công Lạnh
  + Đúc
  + Tôi bề mặt
  + Sản phẩm Hoàn Chỉnh
  + Phương pháp làm mát

A math equation with numbers and symbols

Description automatically generated with medium confidence

* Loại bỏ các biến có chất lượng thấp: Có nhiều hơn 50% giá trị Nah

**A close up of a text

Description automatically generated**

* Loại bỏ các đặc trưng thừa

**A close-up of a computer screen

Description automatically generated**

* Thêm các giá trị thiếu của đặc trưng A5 bằng giá trị trung vị của đặc trưng

**A black text on a white background

Description automatically generated**

* Chuyển các đặc trưng phân loại về dạng số

**A screen shot of a computer code

Description automatically generated**

* Loại bỏ các mẫu trùng lặp

A close-up of a computer

Description automatically generated

A screen shot of a computer code

Description automatically generated

* Áp dụng BorderlineSMOTE để xử lí sự mất cân bằng dữ liệu

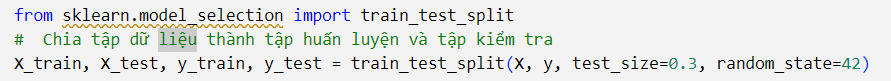
**A computer code with text

Description automatically generated with medium confidence**

* Áp dụng Chuẩn hóa Min-Max với những đặc trưng không phải đặc trưng phân loại

A close-up of a computer code

Description automatically generated

****

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

* Khởi chạy các mô hình

1. **Phân tích hiệu suất của các thuật toán**

Báo cáo sử dụng 2 thuật toán dựa trên ML, cụ thể là SVC và KNN để dự đoán vật liệu được sử dụng. Để kiểm tra hiệu quả của các thuật toán, báo cáo sử dụng bộ dữ liệu gồm 1553 mẫu vật liệu gồm vật liệu được sử dụng và vật liệu không được sử dụng. Các tập dữ liệu này đã cung cấp các điểm chuẩn quan trọng để đánh giá các thuật toán dựa trên ML. Tuy nhiên, chất lượng và tính đa dạng của dữ liệu đào tạo cũng như các tính năng và tham số của thuật toán có thể làm giảm khả năng xác định các ứng dụng độc hại của chúng. Các mô hình dựa trên ML được đề xuất đã chứng minh là một phương pháp hiệu quả để lựa chọn vật liệu được sử dụng.

* 1. **Support Vector Classification - SVC**

A chart of a confused matrix

Description automatically generated with medium confidence A graph with a line

Description automatically generated

A graph of a line graph

Description automatically generated with medium confidence

* + 1. Ma trận nhầm lẫn
    2. Biểu đồ ROC – AUC
    3. Biểu đồ Precison-Recall

Hiệu suất của mô hình SVC

| **Chỉ số** | **Precision** | **Recall** | **F1-score** | **Accuracy** |
| --- | --- | --- | --- | --- |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Giá trị** | 0.829 | 1.000 | 0.906 | 0.899 |

1. **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.
2. **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.
3. **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.
4. **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.

* 1. **K-nearest Neighbor - KNN**

A yellow and purple squares with numbers

Description automatically generated A graph with a line and a blue line

Description automatically generated

A graph of a line

Description automatically generated

* + 1. Ma trận nhầm lẫn
    2. Biểu đồ ROC – AUC
    3. Biểu đồ Precison-Recall

Hiệu suất của mô hình KNN

| **Chỉ số** | **Precision** | **Recall** | **F1-score** | **Accuracy** |
| --- | --- | --- | --- | --- |
| **Giá trị** | 0.932 | 0.974 | 0.952 | 0.952 |

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.

1. **Thảo luận**

Ứng dụng machine learning trong lựa chọn vật liệu sản xuất đã chứng minh khả năng tăng hiệu quả và tiết kiệm thời gian đáng kể. Theo một nghiên cứu gần đây, việc sử dụng ML giúp giảm đến 70% chi phí thử nghiệm và rút ngắn thời gian phát triển sản phẩm từ vài tháng xuống còn vài tuần. ML có thể phân tích nhanh chóng hàng triệu dữ liệu về đặc tính vật liệu như độ bền, độ dẻo, khả năng chịu nhiệt và chi phí, giúp chọn được vật liệu phù hợp một cách tối ưu mà không cần thử nghiệm thủ công trên từng loại. Chẳng hạn, các mô hình ML trong nghiên cứu vật liệu pin đã giảm thiểu thời gian tìm kiếm vật liệu hiệu quả từ 5-10 năm xuống còn khoảng 1-2 năm, đẩy nhanh tốc độ phát triển công nghệ năng lượng tái tạo. Nhờ đó, ML không chỉ giúp các công ty giảm chi phí mà còn nâng cao khả năng cạnh tranh, đồng thời mở ra tiềm năng khám phá các vật liệu mới phục vụ cho ngành sản xuất và công nghệ.

Dưới đây là bảng so sánh hiệu suất của hai mô hình SVC và KNN:

| **Chỉ số** | **SVC** | **KNN** |
| --- | --- | --- |
| **Precision** | 0.829 | 0.932 |
| **Recall** | 1.000 | 0.974 |
| **F1-score** | 0.906 | 0.952 |
| **Accuracy** | 0.899 | 0.952 |
| **AUC (ROC)** | 0.933 | 0.980 |
| **AP (Precision-Recall)** | 0.82 | 0.97 |

**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.

* Để có thể tăng độ chính xác của 2 mô hình, ta có thể sử dụng các cách như
  + Tối ưu hóa tham số
  + Lựa chọn các đặc trưng quan trọng