****

**TRƯỜNG ĐẠI HỌC THỦY LỢI**

KHOA: CÔNG NGHỆ THÔNG TIN

-----🙞🙜🕮🙞🙜-----



**BÁO CÁO THỰC HÀNH BÀI TẬP LỚN**

BỘ MÔN: KHAI PHÁ DỮ LIỆU

**Đề bài:** Sử dụng thuật toán Naïve Bayes và K-Means

để phân loại 3 loại lúa mì dựa vào các thuộc tính của chúng

Giảng viên hướng dẫn: TS. Trần Mạnh Tuấn

Lớp học phần: Khai phá dữ liệu

Sinh viên thực hiện:

Đoàn Hải Long – 1851161680

Đào Văn Hải – 175A071433

Nguyễn Nhật Minh – 175A071346

Nguyễn Quang Minh – 175A071138

Lớp: 60HT – Chuyên ngành hệ thống thông tin

Hà Nội, ngày 10 tháng 11 năm 2021

Mục Lục

[Lời mở đầu 1](#_Toc77089133)

[I. TỔNG QUAN VỀ KHAI KHÁ DỮ LIỆU 2](#_Toc77089134)

[1. Giới thiệu về khai phá dữ liệu (DATA MINING) VÀ KDD 2](#_Toc77089135)

[1.1. Tại sao lại cần khai phá dữ liệu (datamining) 2](#_Toc77089136)

[1.2. Khai phá dữ liệu là gì? 2](#_Toc77089137)

[1.3. Các chức năng chính của khai phá dữ liệu? 6](#_Toc77089138)

[1.4. Ứng dụng của khai phá dữ liệu 7](#_Toc77089139)

[2. Thuật toán áp dụng 7](#_Toc77089140)

[2.1. Thuật toán phân lớp Naïve Bayes (Supervised learning) 7](#_Toc77089141)

[2.2. Thuật toán phân cụm K-means Clustering (Unsupervised learning) 9](#_Toc77089142)

[3. Bài toán đặt ra 11](#_Toc77089143)

[II. KHAI KHÁ DỮ LIỆU 12](#_Toc77089144)

[1. Quá trình khai phá dữ liệu: 12](#_Toc77089145)

[1.1. Từ dữ liệu thô/thực tế: Seeds Data Set: 12](#_Toc77089146)

[1.2. Tiền xử lý dữ liệu 17](#_Toc77089147)

[2. Khai phá dữ liệu 29](#_Toc77089148)

[2.1. Thuật toán phân lớp Naïve Bayes: 29](#_Toc77089149)

[2.2. Thuật toán phân cụm K-Means: 38](#_Toc77089150)

[III. DỰ ÁN 48](#_Toc77089151)

[1. Source Code 48](#_Toc77089152)

[2. Sản phẩm demo 49](#_Toc77089153)

[IV. KẾT LUẬN 55](#_Toc77089154)

[TÀI LIỆU THAM KHẢO 56](#_Toc77089155)

# Lời mở đầu

Sự phát triển của công nghệ thông tin và việc ứng dụng công nghệ thông tin trong nhiều lĩnh vực của đời sống, kinh tế xã hội trong nhiều năm qua cũng đồng nghĩa với lượng dữ liệu đã được các cơ quan thu thập và lưu trữ ngày một tích luỹ nhiều lên. Họ lưu trữ các dữ liệu này vì cho rằng trong nó ẩn chứa những giá trị nhất định nào đó. Tuy nhiên, theo thống kê thì chỉ có một lượng nhỏ của những dữ liệu này (khoảng từ 5% đến 10%) là luôn được phân tích, số còn lại họ không biết sẽ phải làm gì hoặc có thể làm gì với chúng nhưng họ vẫn tiếp tục thu thập rất tốn kém với ý nghĩ lo sợ rằng sẽ có cái gì đó quan trọng đã bị bỏ qua sau này có lúc cần đến nó.

Mặt khác, trong môi trường cạnh tranh, người ta ngày càng cần có nhiều thông tin với tốc độ nhanh để trợ giúp việc ra quyết định và ngày càng có nhiều câu hỏi mang tính chất định tính cần phải trả lời dựa trên một khối lượng dữ liệu khổng lồ đã có. Với những lý do như vậy, các phương pháp quản trị và khai thác cơ sở dữ liệu truyền thống ngày càng không đáp ứng được thực tế đã làm phát triển một khuynh hướng kỹ thuật mới đó là Kỹ thuật phát hiện tri thức và khai phá dữ liệu (KDD - Knowledge Discovery and Data Mining).

Qua môn học “Khai phá dữ liệu” và dưới sự giảng dạy của thầy Trần Mạnh Tuấn đã cung cấp cái nhìn tổng quan về phát hiện tri thức và khai phá dữ liệu. Với những kiến thức đã học, trong bài báo cáo môn học bọn em tập trung vào các thuật toán **phân lớp** và **phân cụm** với bài toán cụ thể là “Sử dụng thuật toán Naïve Bayes và K-Means để phân loại 3 loại lúa mì dựa vào các thuộc tính của chúng”. Bài toán được thực hiện bằng công cụ Weka, bộ mã nguồn mở khá hiệu quả hiện nay.

Nhóm sinh viên thực hiện

# **I. TỔNG QUAN VỀ KHAI KHÁ DỮ LIỆU**

## 1. Giới thiệu về khai phá dữ liệu (DATA MINING) VÀ KDD

### **1.1. Tại sao lại cần khai phá dữ liệu (datamining)**

Khoảng hơn một thập kỷ trở lại đây, lượng thông tin được lưu trữ trên các thiết bị điện tử (đĩa cứng, CD-ROM, băng từ, ...) không ngừng tăng lên.

Sự tích lũy dữ liệu này xảy ra với một tốc độ bùng nổ. Người ta ước đoán rằng lượng thông tin trên toàn cầu tăng gấp đôi sau khoảng hai năm và theo đó số lượng cũng như kích cỡ của các cơ sở dữ liệu (CSDL) cũng tăng lên một cách nhanh chóng.

Nói một cách rõ ràng thì chúng ta đang có quá nhiều dữ liệu thô nhưng lại thiếu tri thức. Liệu chúng ta có thể khai thác được gì lượng dữ liệu thô đó không ?

Data Mining ra đời như một hướng giải quyết hữu hiệu cho câu hỏi vừa đặt ra ở trên. Khá nhiều định nghĩa về Data Mining, tuy nhiên có thể tạm hiểu rằng Data Mining như là một công nghệ tri thức giúp khai thác những thông tin hữu ích từ những kho dữ liệu được tích trữ trong suốt quá trình hoạt động của một công ty, tổ chức nào đó.

### **1.2. Khai phá dữ liệu là gì?**

Khai phá dữ liệu (Data Mining) được định nghĩa như là một quá trình chắt lọc hay khai phá tri thức từ một lượng lớn dữ liệu.

Thuật ngữ Data Mining ám chỉ việc tìm kiếm một tập hợp nhỏ có giá trị từ một số lượng lớn các dữ liệu thô.

Khai phá dữ liệu là một lĩnh vực nhằm tự động khai thác những thông tin tri thức đang tiềm ẩn trong dữ liệu.

Khai phá dữ liệu là một lĩnh vực phát triển bền vững, mang lại nhiều lợi ích, triển vọng, ưu thế hơn hẳn so với các công cụ phân tích dữ liệu truyền thống

Các kỹ thuật được áp dụng dựa trên CSDL, học máy, trí tuệ nhân tạo, lý thuyết thông tin, xác suất thống kê và tính toán hiệu năng cao.

Khai phá tri thức trong CSDL (Knowledge Discovery in Databases - KDD) là mục tiêu chính của Khai phá dữ liệu.

Có nhiều thuật ngữ hiện được dùng cũng có nghĩa tương tự với từ Data Mining như

* Knowledge Mining (khai phá tri thức)
* Knowledge extraction (chắt lọc tri thức)
* Data/patern analysis (phân tích dữ liệu/mẫu)
* Data archaeoloogy (khảo cổ dữ liệu)
* Datadredging (nạo vét dữ liệu), ...

**Định nghĩa:** *Khai phá dữ liệu là một tập hợp các kỹ thuật được sử dụng để tự động khai thác và tìm ra các mối quan hệ lẫn nhau của dữ liệu trong một tập hợp dữ liệu khổng lồ và phức tạp, đồng thời cũng tìm ra các mẫu tiềm ẩn trong tập dữ liệu đó.*

Khai phá dữ liệu là một bước trong bảy bước của quá trình KDD (Knowleadge Discovery in Database) và KDD được xem như 7 quá trình khác nhau theo thứ tự sau:

|  |
| --- |
|  |

1. *Làm sạch dữ liệu (data cleaning & preprocessing): Loại bỏ nhiễu và các dữ liệu không cần thiết.*
2. *Tích hợp dữ liệu: (data integration): quá trình hợp nhất dữ liệu thành những kho dữ liệu (data warehouses & data marts) sau khi đã làm sạch và tiền xử lý (data cleaning & preprocessing).*
3. *Trích chọn dữ liệu (data selection): trích chọn dữ liệu từ những kho dữ liệu và sau đó chuyển đổi về dạng thích hợp cho quá trình khai thác tri thức. Quá trình này bao gồm cả việc xử lý với dữ liệu nhiễu (noisy data), dữ liệu không đầy đủ (incomplete data), …*
4. *Chuyển đổi dữ liệu: Các dữ liệu được chuyển đổi sang các dạng phù hợp cho quá trình xử lý*
5. *Khai phá dữ liệu (data mining): Là một trong các bước quan trọng nhất, trong đó sử dụng những phương pháp thông minh để chắt lọc ra những mẫu dữ liệu.*
6. *Ước lượng mẫu (knowledge evaluation): Quá trình đánh giá các kết quả tìm được thông qua các độ đo nào đó.*
7. *Biểu diễn tri thức (knowledge presentation): Quá trình này sử dụng các kỹ thuật để biểu diễn và thể hiện trực quan cho người dùng.*

|  |
| --- |
|  |

**Các kỹ thuật áp dụng trong khai phá dữ liệu:**

* Học có giám sát (Supervised learning): Quá trình gán nhãn lớp cho các phần tử trong cơ sở dữ liệu dựa trên một tập các ví dụ huấn luyện và các thông tin về nhãn lớp đã biết.

|  |
| --- |
|  |

* Học không có giám sát (Unsupervised learning): Quá trình phân chia một tập dữ liệu thành các cụm (clustering) dữ liệu tương tự nhau mà chưa biết trước các thông tin về lớp/tập các ví dụ huấn luyện.
* Học nửa giám sát (Semi - Supervised learning): Là quá trình phân chia một tập dữ liệu thành các lớp dựa trên một tập nhỏ các ví dụ huấn luyện và một số các thông tin về một số nhãn lớp đã biết trước.

### **1.3. Các chức năng chính của khai phá dữ liệu?**

Data Mining được chia nhỏ thành một số hướng chính như sau:

* **Mô tả khái niệm (concept description)**: thiên về mô tả, tổng hợp và tóm tắt khái niệm. Ví dụ: tóm tắt văn bản, báo cáo, ...
* **Luật kết hợp (association rules)**: là dạng luật biểu diễn tri thứ ở dạng khá đơn giản. Ví dụ: “70% người nếu mua máy tính thì có tới 90% trong số họ sẽ mua thêm chuột và bán phím”. Luật kết hợp được ứng dụng nhiều trong lĩnh vực kính doanh, y học, tin-sinh, tài chính & thị trường chứng khoán, ...
* **Phân lớp và dự đoán (classification & prediction)**: xếp một đối tượng vào một trong những lớp đã biết trước. Ví dụ: phân lớp vùng địa lý theo dữ liệu thời tiết. Hướng tiếp cận này thường sử dụng một số kỹ thuật của machine learning như cây quyết định (decision tree), mạng nơ ron nhân tạo (neural network), ... Người ta còn gọi **phân lớp là học có giám sát (học có thầy)**.
* **Phân cụm (clustering)**: xếp các đối tượng theo từng cụm (số lượng cũng như tên của cụm chưa được biết trước. Người ta còn gọi **phân cụm là học không giám sát (học không thầy)**.
* **Khai phá chuỗi (sequential/temporal patterns)**: tương tự như khai phá luật kết hợp nhưng có thêm tính thứ tự và tính thời gian. Hướng tiếp cận này được ứng dụng nhiều trong lĩnh vực tài chính và thị trường chứng khoán vì nó có tính dự báo cáo.

### **1.4. Ứng dụng của khai phá dữ liệu**

Data Mining tuy là một hướng tiếp cận mới nhưng thu hút được rất nhiều sự

quan tâm của các nhà nghiên cứu và phát triển nhờ vào những ứng dụng thực tiễn của

nó. Chúng ta có thể liệt kê ra đây một số ứng dụng điển hình:

* Phân tích dữ liệu và hỗ trợ ra quyết định (data analysis & decision support)
* Điều trị y học (medical treatment)
* Text mining & Web mining
* Tin-sinh (bio-informatics)
* Tài chính và thị trường chứng khoán (finance & stock market)
* Bảo hiểm (insurance)
* Nhận dạng (pattern recognition)
* …

## 2. Thuật toán áp dụng

### **2.1. Thuật toán phân lớp Naïve Bayes (Supervised learning)**

**2.1.1. Khái niệm**

**Naïve Bayes Classification (NBC)** là một thuật toán thuộc loại **học có giám sát** dựa trên định lý Bayes về lý thuyết xác suất để đưa ra các phán đoán cũng như phân loại dữ liệu dựa trên các dữ liệu được quan sát và thống kê.

Naïve Bayes Classification là một trong những thuật toán được ứng dụng rất nhiều trong các lĩnh vực Machine learning dùng để đưa các dự đoán chính xác nhất dự trên một tập dữ liệu đã được thu thập, vì nó khá dễ hiểu và độ chính xác cao.

Nó thuộc vào nhóm **Supervised Machine Learning Algorithms** (thuật toán học có hướng dẫn), tức là máy học từ các ví dụ từ các mẫu dữ liệu đã có.

**2.1.2. Thuật toán phân lớp Naïve Bayes**

Theo định lý Bayes, ta có công thức tính xác suất ngẫu nhiên của sự kiện y khi biết x như sau:



Giả sử ta phân chia 1 sự kiện x thành nn thành phần khác nhau , ,…, ​  
Naive Bayes theo đúng như tên gọi dựa vào một giả thiết ngây thơ rằng , ,…, ​ là các thành phần độc lập với nhau. Từ đó ta có thể tính được:



Do đó ta có:



∝ là phép tỉ lệ thuận.

Trên thực tế thì ít khi tìm được dữ liệu mà các thành phần là hoàn toàn độc lập với nhau. Tuy nhiên giả thiết này giúp cách tính toán trở nên đơn giản, training data nhanh, đem lại hiệu quả bất ngờ với các lớp bài toán nhất định.

Cách xác định các thành phần (class) của dữ liệu dựa trên giả thiết này có tên là Naive Bayes Classifier.

Bộ phân lớp Naive bayes (simple byes classifier) hoạt động như sau:

* Gọi D là tập dữ liệu huấn luyện, trong đó mỗi phần tử dữ liệu X được biểu diễn bằng một vector chứa n giá trị thuộc tính , , ...,  = {, , ..., }
* Giả sử có m lớp , , ..., . Cho một phần tử dữ liệu X, bộ phân lớp sẽ gán nhãn cho X là lớp có xác suất hậu nghiệm lớn nhất. Cụ thể, bộ phân lớp Bayes sẽ dự đoán X thuộc vào lớp  nếu và chỉ nếu:



* Giá trị này sẽ tính dựa trên định lý Bayes.
* Để tìm xác suất lớn nhất, ta nhận thấy các giá trị P(X) là giống nhau với mọi lớp nên không cần tính. Do đó ta chỉ cần tìm giá trị lớn nhất của P(X|) \* P()).
* Chú ý rằng P()) được ước lượng bằng || / |D|, trong đó  là tập các phần tử dữ liệu thuộc lớp . Nếu xác suất tiền nghiệm P() cũng không xác định được thì ta coi chúng bằng nhau P() = P() = ... = P(), khi đó ta chỉ cần tìm giá trị P(X|) lớn nhất.
* Khi số lượng các thuộc tính mô tả dữ liệu là lớn thì chi phí tính toàn P(X|) là rất lớn, dó đó có thể giảm độ phức tạp của thuật toán Naive Bayes giả thiết các thuộc tính độc lập nhau. Khi đó ta có thể tính:



### **2.2. Thuật toán phân cụm K-means Clustering (Unsupervised learning)**

**2.2.1. Khái niệm**

**K-means** **Clustering** là một thuật toán phân cụm đơn giản thuộc loại học **không giám sát**(tức là dữ liệu không có nhãn) và được sử dụng để giải quyết bài toán phân cụm.

Ý tưởng của thuật toán phân cụm k-means là phân chia 1 bộ dữ liệu thành các cụm khác nhau. Trong đó số lượng cụm được cho trước là k. Công việc phân cụm được xác lập dựa trên nguyên lý: Các điểm dữ liệu trong cùng 1 cụm thì phải có cùng 1 số tính chất nhất định.

Tức là giữa các điểm trong cùng 1 cụm phải có sự liên quan lẫn nhau. Đối với máy tính thì các điểm trong 1 cụm đó sẽ là các điểm dữ liệu gần nhau.

Thuật toán phân cụm K-Means thường được sử dụng trong các ứng dụng cỗ máy tìm kiếm, phân đoạn khách hàng, thống kê dữ liệu,…

**2.2.2. Thuật toán phân cụm K-Means**

Các tâm cụm cực tiểu sự biến đổi giữa các cụm



Các tâm cụm (trung tâm của cụm):



Bài toán cực tiểu hóa này là tối ưu tổ hợp: Giải pháp cho cực tiểu hóa địa phương ta sử dụng phương pháp lặp

Công thức tính khoảng cách các điểm đến tâm



* Input
* Tập mẫu X = { | i = 1, 2, ..., N},
* Số cụm: K
* Output
* Các cụm (k = 1 ÷ K) tách rời và hàm mục tiêu J đạt cực tiểu

Bộ phân cụm K-Means hoạt động như sau:

* Khởi tạo: Chọn ngẫu nhiên K tâm cụm
* Tính toán khoảng cách từ các đối tượng đến các tâm để phân hoạch dữ liệu (bằng cách gán mỗi đối tượng vào cụm mà nó gần tâm nhất)
* Tính lại các tâm cụm mới trong mỗi cụm
* Lặp lại 2 và 3 cho đến khi “thỏa mãn điều kiện” (khi các tâm cụm ổn định và các đối tượng không dịch chuyển giữa các cụm)

## 3. Bài toán đặt ra

Sử dụng thuật toán phân lớp Naïve Bayes và phân cụm K-Means để xác định loại giống lúa mì trong các giống sau: Kama, Rosa và Canadian.

# **II. KHAI KHÁ DỮ LIỆU**

## 1. Quá trình khai phá dữ liệu:

### **1.1. Từ dữ liệu thô/thực tế: Seeds Data Set:**

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

**Data Set Information:**

Nhóm được kiểm tra bao gồm các hạt nhân thuộc ba loại lúa mì khác nhau: Kama, Rosa và Canada, mỗi loại 70 nguyên tố, được chọn ngẫu nhiên cho thí nghiệm. Các hình dung chất lượng cao của cấu trúc nhân bên trong được phát hiện bằng kỹ thuật X-quang mềm.

Nó không phá hủy và rẻ hơn đáng kể so với các kỹ thuật hình ảnh phức tạp khác như kính hiển vi quét hoặc công nghệ laser. Hình ảnh được ghi lại trên các tấm X-ray KODAK 13x18 cm.

Các nghiên cứu được thực hiện bằng cách sử dụng hạt lúa mì thu hoạch kết hợp có nguồn gốc từ các cánh đồng thí nghiệm, được khám phá tại Viện Vật lý học của Viện Hàn lâm Khoa học Ba Lan ở Lublin.

Tập dữ liệu có thể được sử dụng cho các nhiệm vụ phân loại và phân tích cụm.

**Attribute Information:**

Để xây dựng dữ liệu, người ta đo bảy thông số hình học của hạt lúa mì:

1. diện tích A
2. chu vi P
3. độ nén C = 4 \* pi \* A / P ^ 2
4. chiều dài của nhân
5. chiều rộng của nhân
6. hệ số không đối xứng
7. chiều dài của rãnh nhân.

Tất cả các tham số này đều có giá trị thực liên tục.

### **1.2. Tiền xử lý dữ liệu**

**1.2.1. Tiền xử lý dữ liệu**

**Các kỹ thuật tiền xử lý dữ liệu**

* Làm sạch dữ liệu (data cleaning/cleansing)
  + Tóm tắt hoá dữ liệu: nhận diện đặc điểm chung của dữ liệu và sự hiện diện của nhiễu hoặc các phần tử kì dị (outliers)
  + Xử lý dữ liệu bị thiếu (missing data)
  + Xử lý dữ liệu bị nhiễu (noisy data)
* Tích hợp dữ liệu (data integration)
  + Tích hợp lược đồ (schema integration) và so trùng đối tượng (object matching)
  + Vẫn đề dư thừa (redundancy)
  + Phát hiện và xử lý mâu thuẫn giá trị dữ liệu (detection and resolution of data value conflicts)
* Biến đổi dữ liệu (data transformation)
  + Làm trơn dữ liệu (smoothing)
  + Kết hợp dữ liệu (aggregation)
  + Tổng quát hóa dữ liệu (generalization)
  + Chuẩn hóa dữ liệu (normalization)
  + Xây dựng thuộc tính (attribute/feature construction)
* Thu giảm dữ liệu (data reduction)
  + Kết hợp khối dữ liệu (data cube aggregation)
  + Chọn tập con các thuộc tính (attribute subset selection)
  + Thu giảm chiều (dimensionality reduction)
  + Thu giảm lượng (numerosity reduction)
  + Tạo phân cấp ý niệm (concept hierarchy generation) và rời rạc hóa (discretization)

**1.2.2. Làm sạch dữ liệu**

Là quá trình nhận dạng dữ liệu đã có để tiến hành xử lý các dữ liệu bị thiếu (missing data) xử lý dữ liệu bị nhiễu (noisy data) và không nhất quán.

Dữ liệu ban đầu:

|  |
| --- |
|  |

Copy dữ liệu vào Excel và lưu file dưới dạng file .csv:

|  |
| --- |
|  |

Open file Seed\_Data\_BTL.csv vào weka

|  |
| --- |
|  |

Nhấm Edit để làm sạch dữ liệu:

* Nếu cột dữ liệu nào thiếu thì có thể bổ sung dữ liệu đó bằng giá trị Mean thông qua tính năng Set missing values to…

|  |
| --- |
|  |

Chạy MathExpression để biến tất cả các giá trị thành những giá trị rời rạc

|  |
| --- |
| weka.filters.unsupervised.attribute.MathExpression |

* Chọn Choose **➞** Unsupervised **➞** Attribute **➞** MathExpression
* Chọn các thuộc tính thỏa mãn hình vẽ **➞** Chọn OK
* Chọn Apply

|  |
| --- |
|  |

Chuyển thuộc tính Type từ dạng Numeric sang Nominal để phù hợp với việc phân lớp:

|  |
| --- |
| weka.filters.unsupervised.attribute.NumericToNominal |

* Chọn Choose **➞** Unsupervised **➞** Attribute **➞** NumericToNominal
* Chọn AttributeIndices là last hoặc 8 **➞** Chọn OK
* Chọn Apply

|  |
| --- |
|  |

Các thuộc tính và dữ liệu sau khi Apply:

|  |
| --- |
|  |

Kiểm tra dữ liệu có phần tử ngoại lai bằng phương thức interquartile range:

|  |
| --- |
| weka.filters.unsupervised.attribute.InterquatileRange |

* Chọn Choose **➞** Unsupervised **➞** Attribute **➞** InterquatileRange
* Chọn Apply

|  |
| --- |
|  |

Kết quả sau khi chạy:

|  |
| --- |
|  |

Giảm kích thước dữ liệu bằng việc loại bỏ những thuộc tính dư thừa với phương thức AttributeSelection:

|  |
| --- |
| weka.filters.supervised.attribute.AttributeSeclection |

* Chọn Choose **➞** Supervised **➞** Attribute **➞** AttributeSelection
* Chọn Apply

|  |
| --- |
|  |

Kết quả sau khi chạy:

* Giá trị Attributes từ 10 giảm xuống còn 7 (đã lại bỏ những thuộc tính dư thừa)

|  |
| --- |
|  |

## 2. Khai phá dữ liệu

### **2.1. Thuật toán phân lớp Naïve Bayes:**

|  |
| --- |
|  |

***Phương pháp 1: Dùng test option Percentage Split***

**Percentage Split:**

* Cho biết tỉ lệ phân chia là bao nhiêu % thì đạt hiệu quả phân lớp cao nhất.
* Tỷ lệ phân chia cho biết số mẫu được chọn cho tập huấn luyện và số mẫu được chọn cho tập test.

**Ứng dụng bài toán:**

* Với 70% là training set, 30% còn lại là test set

**Kết quả chạy:**

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

**Giải thích kết quả:**

Classifier model: Naive Bayes Classifier

**Percentage Split:**

* Trường hợp phân lớp chính xác: 58 (92.0635%)
* Trường hợp phân lớp không chính xác: 5 (7.9365%)
* Sai số tuyệt đối trung bình: 0.052
* Sai số bình phương trung bình: 0.2092
* Sai số tuyệt đối tương đối: 11.6088 %
* Sai số bình phương tương đối: 43.9298 %
* Tổng số trường hợp phân lớp: 63
* Ma trận hỗn loạn

|  |
| --- |
|  |

* Hàng 1, có 14 trường hợp phân lớp đúng (lớp a) và 1 trường hợp phân lớp sai vào lớp c.
* Hàng 2, có 21 trường hợp phân lớp đúng (lớp b) và 3 trường hợp phân lớp sai vào lớp a.
* Hàng 1, có 23 trường hợp phân lớp đúng (lớp c) và 1 trường hợp phân lớp sai vào lớp a.

***Phương pháp 2: Dùng test option 10-fold cross-validation:***

**Cross-validation:**

* Một kĩ thuật được sử dụng để kiểm tra hiệu xuất của mô hình tiên đoán. Cross-validation tiên đoán sự ăn khớp của một mô hình tới một tập giả thiết hợp lệ (test set) khi mà ta không có sẵn tập hợp lệ tường minh.
* k-folds **cross-validation:** Tập mẫu ban đầu được phân chia ngẫu nhiên tới k tập con. Với k tập mẫu con này, một mẫu đơn được dùng như dữ liệu đánh giá cho việc kiểm tra mô hình, và k – 1 tập mẫu còn lại được sử dụng như dữ liệu training. Tiến hành đánh giá chéo được lặp lại k lần (tham sô Folds trong weka). Lấy trung bình k kết quả thu được ra có một đánh giá cho mô hình.

**Ứng dụng bài toán:**

* Weka lấy 210 dữ liệu được gắn nhãn.
* Nó tạo ra 10 bộ có kích thước bằng nhau. Mỗi tập hợp được chia thành hai nhóm: 200 dữ liệu có nhãn được sử dụng để đào tạo và 10 dữ liệu có nhãn được sử dụng để kiểm tra.
* Nó tạo ra một bộ phân loại với một thuật toán từ 200 dữ liệu được gắn nhãn và áp dụng nó trên 10 dữ liệu thử nghiệm cho tập 1.
* Nó thực hiện điều tương tự cho tập 2 đến 10 và tạo ra 9 bộ phân loại khác.
* Nó tính **trung bình hiệu suất** của 10 bộ phân loại được tạo ra từ 10 bằng nhau có kích thước (200 bộ đào tạo và 10 bộ kiểm tra)

|  |
| --- |
|  |

**Kết quả chạy:**

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

**Giải thích kết quả:**

Classifier model: Naive Bayes Classifier

**Cross-validation:**

* Trường hợp phân lớp chính xác: 191 (90.9524%)
* Trường hợp phân lớp không chính xác: 19 (9.0476%)
* Sai số tuyệt đối trung bình: 0.0612
* Sai số bình phương trung bình: 0.2154
* Sai số tuyệt đối tương đối: 13.7722 %
* Sai số bình phương tương đối: 45.693 %
* Tổng số trường hợp phân lớp: 210
* Ma trận hỗn loạn

|  |
| --- |
|  |

* Hàng 1, có 60 trường hợp phân lớp đúng (lớp a) và 3 trường hợp phân lớp sai vào lớp b, 7 trường hợp phân lớp sai vào lớp c.
* Hàng 2, có 66 trường hợp phân lớp đúng (lớp b) và 4 trường hợp phân lớp sai vào lớp a.
* Hàng 1, có 65 trường hợp phân lớp đúng (lớp c) và 5 trường hợp phân lớp sai vào lớp a.

Trong đó:

Chi tiết độ chính xác theo các lớp:

* **Tỷ lệ TP** : Tỷ lệ dương tính thực (các trường hợp được phân loại chính xác như một lớp đã cho)
* **Tỷ lệ FP** : Tỷ lệ dương tính giả (các trường hợp được phân loại sai là một lớp nhất định)
* **Độ chính xác** : Tỷ lệ các trường hợp thực sự là một lớp chia cho tổng số các trường hợp được phân loại là lớp đó
* **Recall**: Tỷ lệ các trường hợp được phân loại là một lớp nhất định chia cho tổng số thực tế trong lớp đó (tương đương với tỷ lệ TP)
* **F-Measure** : Một biện pháp kết hợp cho độ chính xác và thu hồi được tính bằng 2 \* Precision \* Recall / (Precision + Recall).

### **2.2. Thuật toán phân cụm K-Means:**

|  |
| --- |
|  |

* Chọn mode Cluster:
* Loại bỏ thuộc tính Type: **Click Ignore Attributes**

|  |
| --- |
|  |

* Chia số cụm theo số type: **numCluters: 3**

|  |
| --- |
|  |

***Phương pháp 1: Dùng test option Percentage Split***

**Percentage Split:**

* Cho biết tỉ lệ phân chia là bao nhiêu % thì đạt hiệu quả phân cụm cao nhất.
* Tỷ lệ phân chia cho biết số mẫu được chọn cho tập huấn luyện và số mẫu được chọn cho tập test.

**Ứng dụng bài toán:**

* Với 70% là training set, 30% còn lại là test set

|  |
| --- |
|  |

**Kết quả chạy:**

|  |
| --- |
|  |

**Phân cụm với tập training:**

|  |
| --- |
|  |

**Giải thích kết quả:**

* Số lần lặp tính tâm cụm: 7
* Tổng số lỗi bình phương trong cụm: 16.566515670437447
* Điểm khởi tạo tâm cụm đầu tiên được chọn ngẫu nhiên:

|  |
| --- |
|  |

* Tâm cụm cuối cùng được tính:

|  |
| --- |
|  |

**Phân cụm với tập test:**

|  |
| --- |
|  |

**Giải thích kết quả:**

* Số lần lặp tính tâm cụm: 5
* Tổng số lỗi bình phương trong cụm: 12.899780330832126
* Điểm khởi tạo tâm cụm đầu tiên được chọn ngẫu nhiên:

|  |
| --- |
|  |

* Tâm cụm cuối cùng được tính:

|  |
| --- |
|  |

* Số trường hợp được phân cụm

|  |
| --- |
|  |

* 21 trường hợp được phân vào cụm 0 (33%)
* 16 trường hợp được phân vào cụm 1 (25%)
* 26 trường hợp được phân vào cụm 2 (41%)

***Phương pháp 2: Dùng tập huấn luyện để kiểm thử***

**Use Training Set:**

**Ứng dụng bài toán:**

**Kết quả chạy:**

|  |
| --- |
|  |

|  |
| --- |
|  |

**Giải thích kết quả:**

Phân cụm với tập huấn luyện:

* Số lần lặp tính tâm cụm: 7
* Tâm cụm đầu tiên được chọn ngẫu nhiên như hình trên

|  |
| --- |
|  |

* Tâm cụm cuối cùng được tính như hình trên

|  |
| --- |
|  |

* Số trường hợp được phân cụm:
* 62 trường hợp được phân vào cụm 0 (30%)
* 85 trường hợp được phân vào cụm 1 (40%)
* 63 trường hợp được phân vào cụm 2 (30%)

# **III. DỰ ÁN**

## 1. Source Code

* Import 1 số thư viện cần dùng

|  |
| --- |
| import streamlit as st  import pandas as pd  import numpy as np  import matplotlib.pyplot as plt  import seaborn as sns  from sklearn.model\_selection import train\_test\_split  from sklearn import metrics  from sklearn.preprocessing import LabelEncoder  from sklearn.metrics import classification\_report  from sklearn.preprocessing import StandardScaler  from sklearn.naive\_bayes import GaussianNB  from sklearn.cluster import KMeans |

* Thuật toán Naïve Bayes

|  |
| --- |
| # Training Naive-bayes  @st.cache(suppress\_st\_warning=True)  def Naive\_bayes(X\_train, X\_test, y\_train, y\_test):  gnb = GaussianNB()  gnb\_model = gnb.fit(X\_train, y\_train)  y\_pred = gnb\_model.predict(X\_test)  score = metrics.accuracy\_score(y\_test, y\_pred) \* 100  report = classification\_report(y\_test, y\_pred)  return score, report, gnb |

## 2. Sản phẩm demo

* Dữ liệu

|  |
| --- |
|  |

* Tóm tắt dữ liệu

|  |
| --- |
|  |

* Biểu đồ phân bố hạt giống theo chiều dài của nhân

|  |
| --- |
|  |

* Biểu đồ phân bố hạt giống theo chiều rộng của nhân

|  |
| --- |
|  |

* Biểu đồ các yếu tố khác

|  |
| --- |
|  |

* Biểu đồ phân phối các biến số

|  |
| --- |
|  |

* Dự đoán kết quả sử dụng thuật toán Naïve Bayes

|  |
| --- |
|  |

|  |
| --- |
|  |

# **IV. KẾT LUẬN**

Phân cụm và phân lớp là 2 lĩnh vực khá quan trọng trong khai phá dữ liệu xu hướng trong tương lai, nó được ứng dụng trong nhiều ngành như y tế, thương mại... Hoàn thành đề tài “Sử dụng thuật toán naive bayes và k-means để phân loại 3 loại lúa mì dựa vào các thuộc tính của chúng”. Nhóm em đã đạt được một số kết quả như sau:

* Tìm hiểu tổng quan về khai phá dự liệu, bài toán phân lớp, phân cụm, sử dụng thuật toán naive bayes và k-means để phân loại 3 loại lúa mì dựa vào các thuộc tính của chúng.
* Thu thập dữ liệu, tiền xử lý dữ liệu bằng excel và weka. Xây dựng nên mô hình phân lớp và phân cụm trên weka.

Tuy nhiên bài tập nhóm vẫn còn một số hạn chế:

* Kết quả dự đoán tương đối cao nhưng vẫn chưa được tốt nhất.

Trong quá trình hoàn thành bài tập lớn, nhóm em đã cố gắng tìm hiểu và tham khảo các tài liệu liên quan. Tuy nhiên, thời gian có hạn nên chúng em sẽ không tránh khỏi những thiếu sót, rất mong nhận được sự đóng góp ý kiến của quý thầy cô và các bạn để báo cáo và kỹ năng của chúng em ngày được hoàn thiện hơn và có thế áp dụng được trong thực tiễn.

# **TÀI LIỆU THAM KHẢO**

**1. Tài liệu tham khảo**

[1] TS.Trần Mạnh Tuấn

<https://sites.google.com/site/tranmanhtuantlu/cac-mon-day/khai-pha-du-lieu?fbclid=IwAR2gbOp2kIOOcNROqCK7wQcJsW44OY1dcyzkm-pSy9lEE3jYfARRMOQ7jYo>

[2] Machine Learning cơ bản

<https://machinelearningcoban.com/2016/12/26/introduce/>

[3] Streamlit

<https://streamlit.io/>

**2. Dữ liệu**

[1] Dữ liệu tham khảo

<https://archive.ics.uci.edu/ml/datasets/seeds>

[2] Dữ liệu Kaggle

<https://www.kaggle.com/dongeorge/seed-from-uci>