***Mục Lục***

[***1.*** ***Phân lớp:*** 2](#_Toc87110789)

[***1.1.*** ***Cây quyết định:*** 2](#_Toc87110790)

[1.1.1. Mô tả dữ liệu 2](#_Toc87110791)

[1.1.2. Trình bày ý tưởng thuật toán 3](#_Toc87110792)

[1.1.3. Triển khai mô hình 4](#_Toc87110793)

[1.1.4. Áp dụng mô hình 8](#_Toc87110794)

[1.1.5. Nhận xét 10](#_Toc87110795)

[***1.2.*** ***K- NN*** 10](#_Toc87110796)

[1.2.1. Mô tả dữ liệu 10](#_Toc87110797)

[1.2.2. Trình bày ý tưởng thuật toán 10](#_Toc87110798)

[1.2.3. Triển khai mô hình 11](#_Toc87110799)

[1.2.4. Áp dụng mô hình 14](#_Toc87110800)

[1.2.5. Nhận xét 14](#_Toc87110801)

# ***Phân lớp:***

## ***Cây quyết định:***

### Mô tả dữ liệu

* Lựa chọn dữ liệu: Lấy dữ liệu dataset tại link: <https://archive.ics.uci.edu/ml/datasets/Flags>
* Giới thiệu dữ liệu:
* Tệp dữ liệu này chứa thông tin chi tiết về các quốc gia khác nhau và cờ của họ. Trong tệp này, các trường được phân tách bằng dấu cách (không phải dấu phẩy).
* Dữ liệu gồm 10 thuộc tính có giá trị số. Phần còn lại là Boolean- hoặc giá trị danh nghĩa.
* Thông tin các thuộc tính:

1. Tên: Tên quốc gia liên quan  
2. Diện tích đất: 1 = N.America, 2 = S.America, 3 = Châu Âu, 4 = Châu Phi, 4 = Châu Á, 6 = Châu Đại Dương  
3. Khu vực: Góc phần tư địa lý, dựa trên Greenwich và Xích đạo; 1 = NE, 2 = SE, 3 = SW, 4 = NW  
4. Diện tích: tính bằng nghìn km vuông  
5. Dân số: tính bằng triệu  
6. Ngôn ngữ: 1 = tiếng Anh, 2 = tiếng Tây Ban Nha, 3 = tiếng Pháp, 4 = tiếng Đức , 5 = Slavic, 6 = Ấn-Âu khác, 7 = Trung Quốc, 8 = Ả Rập, 9 = Nhật / Thổ Nhĩ Kỳ / Phần Lan / Magyar, 10 = Khác  
7. tôn giáo: 0 = Công giáo, 1 = Cơ đốc giáo khác, 2 = Hồi giáo, 3 = Phật giáo, 4 = Hindu, 5 = Dân tộc, 6 = Người theo chủ nghĩa Mác, 7 = Người khác  
8. Thanh: Số thanh dọc trên lá cờ  
9. Sọc: Số sọc ngang trên lá cờ  
10. Màu sắc: Số màu khác nhau trong lá cờ  
11. Đỏ: 0 nếu không có màu đỏ, 1 nếu màu đỏ xuất hiện trên lá cờ  
12. Xanh lá cây: giống với màu xanh lá cây  
13. Xanh lam: giống với màu xanh lam  
14. Vàng: giống với vàng (cũng có màu vàng)  
15. Trắng: giống với màu trắng  
16. Màu đen: giống màu đen  
17. Màu cam: giống màu da cam (cũng có màu nâu)  
18. Màu mainhue: màu chủ đạo trong cờ (hòa-giải quyết định bằng cách lấy màu trên cùng, nếu không đạt thì đó là màu trung tâm nhất, và nếu không đạt thì sắc độ ngoài cùng bên trái)  
19. Vòng tròn: Số vòng tròn trên lá cờ  
20. Dấu chéo: Số đường chéo (thẳng đứng)  
21. Độ đậm của đường chéo: Số đường chéo  
22. phần tư: Số phần chia nhỏ  
23. Biểu tượng: Số biểu tượng mặt trời hoặc ngôi sao  
24. Lưỡi liềm: 1 nếu có biểu tượng trăng lưỡi liềm, khác 0  
25. Hình tam giác: 1 nếu có hình tam giác, 0 nếu không  
26. Biểu tượng: 1 nếu có hình ảnh vô tri vô giác (ví dụ: một chiếc thuyền), nếu không thì 0  
27. Hình ảnh động: 1 nếu có hình ảnh động (ví dụ: đại bàng, cây cối, bàn tay người), 0 nếu không  
28. Văn bản: 1 nếu có bất kỳ chữ cái hoặc chữ viết nào trên lá cờ (ví dụ: khẩu hiệu hoặc khẩu hiệu), 0 nếu không thì  
29. Màu góc trên bên trái: Màu ở góc trên bên trái (di chuyển sang phải để quyết định hòa)  
30. Màu góc dưới bên trái: Tô màu ở góc dưới bên trái (di chuyển sang trái để quyết định hòa)

* Mục đích: Với dữ liệu này, tôi có thể thử những thứ như dự đoán tôn giáo của một quốc gia từ kích thước và màu sắc trên lá cờ của quốc gia đó.Label được chọc để dự đoán : Label Religion (Tôn giáo).

### Trình bày ý tưởng thuật toán

* Cây quyết định là thuật toán được xây dựng trên mô hình một cây gồm các mục đích khác nhau để thể hiện cấu trúc của một hệ thống ra quyết định, hay nói cách khác là cách con người tư duy logic để đi đến quyết định cuối cùng. Cây quyết định là một cây gồm các nút (node) biểu diễn một đặc điểm (thuộc tính), mỗi liên kết nhánh (branch) thể hiện một quy luật (rule) và mỗi lá cho thấy một kết quả tương ứng (có thể là một giá trị cụ thể hoặc một nhánh tiếp tục).

Diagram

Description automatically generated

* Các thành phần chính của cây quyết định:
  + Gốc (Root): Điểm ngọn chứa giá trị của biến đầu tiên dùng để phân nhánh
  + Nút (Node): Các điểm bên trong thân cây chứa các thuộc tính, giá trị dữ liệu được dùng để xét cho các phân nhánh tương ứng
  + Lá (Leaf): Là điểm chứa giá trị của biến phân loại sau cùng
  + Nhánh (Branch): là quy luật phân nhánh, hay nói đơn giản là mối quan hệ giữa các giá trị của nút với lá hoặc giữa các nút kế tiếp nhau.
* Các bước thực hiện:
  + Bước 1: Xác định các thuộc tính cần phân nhánh và tần suất xuất hiện của các giá trị thuộc tính ứng với từng giá trị của biến mục tiêu.
  + Bước 2: Áp dụng công thức GINI Index và GINI Split cho từng thuộc tính để tìm ra cách phân nhánh tốt nhất.
  + Bước 3: Phân các đối tượng vào nhánh tương ứng.
* Công thức hệ số GINI (tính độ đồng nhất của 1 node): Gini đo lường sự bất bình đẳng giữa các giá trị phân phối.
* GINI = 0 thể hiện sự bình đẳng hoàn hảo.
* GINI = 1 thể hiện sự bất bình đẳng tối đa.

A picture containing text

Description automatically generated

* Công thức GINI phân nhánh (tìm cách phân nhánh tối ưu): Khi 1 nút p được chia thành k phân vùng (children) thì GINI của phân chia được tính như sau:

Text

Description automatically generated with medium confidence

* Trong đó, 𝑛𝑖 là số điểm dữ liệu trong node i, còn n là tổng số điểm dữ liệu. Hệ số GINI split càng nhỏ tức cách phân nhánh càng tối ưu.

### Triển khai mô hình

* Tệp dữ liệu cho cây quyết định:

A picture containing text, wall, indoor, white

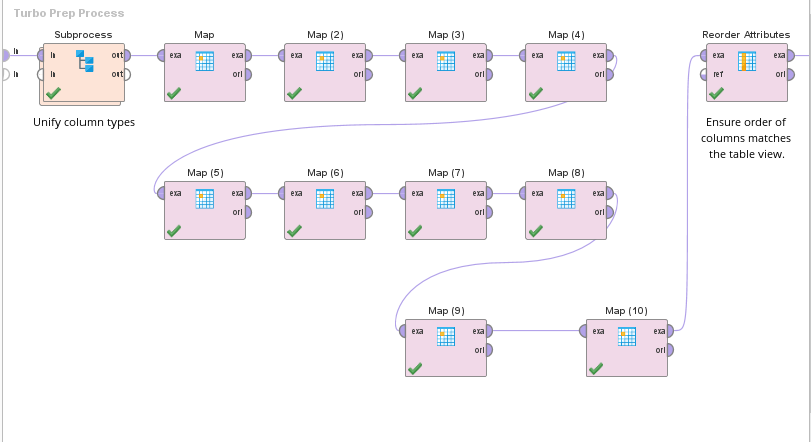
Description automatically generated

* Process triển khai mô hình:

Diagram

Description automatically generated

* Read Excel: Toán tử giúp đọc một tập hợp ví dụ từ các tệp bảng tính Excel.
* Numerical to polynominal: Toán tử Numerical to Polynominal được sử dụng để thay đổi kiểu thuộc tính số thành kiểu polynominal, sau đó nó ánh xạ tất cả các giá trị số đơn giản đến các giá trị danh nghĩa tương ứng.
* Turbo Prep Process: Toán tử này giới thiệu một quy trình trong một quy trình. Bất cứ khi nào một toán tử Quy trình con được tiếp cận trong quá trình thực thi quy trình, trước tiên toàn bộ quy trình con sẽ được thực thi. Khi quá trình thực thi quy trình con hoàn tất, luồng được trả về quy trình (quy trình mẹ). Một quy trình con có thể được coi là một đơn vị nhỏ của một quy trình, giống như trong quy trình, tất cả các toán tử và tổ hợp các toán tử có thể được áp dụng trong một quy trình con. Đó là lý do tại sao một quy trình con cũng có thể được định nghĩa là một chuỗi các toán tử được áp dụng sau đó.



* + Trong Turbor Prep Process : Thực hiện mapping các giá trị tương ứng theo dữ liệu đang áp dụng.
  + Map : Toán tử này ánh xạ các giá trị được chỉ định của các thuộc tính đã chọn thành các giá trị mới. Toán tử này có thể được áp dụng trên cả thuộc tính số và thuộc tính danh nghĩa. ( Map các trường giá trị Lanmass, Zone, Language…)
* Set Role: Toán tử này được sử dụng để thay đổi vai trò của một hoặc nhiều thuộc tính.Trong bài toán này chọn thuộc tính Religion để phân lớp.
* Split Data: Toán tử này tạo ra số lượng các tập con mong muốn của tập mẫu đã cho. Trong bài toán này tạo ra 2 tập con : 80% dữ liệu cho training và 20% dữ liệu cho testing.
* Cross Validation: Toán tử này thực hiện xác nhận chéo để ước tính hiệu suất thống kê của một mô hình học tập.

Diagram

Description automatically generated with low confidence

* Apply model : công cụ giúp áp dụng mô hình thuật toán cây quyết định với dữ liệu testing để dự đoán kết quả của Label.
* Performance: công cụ thực hiện tính hiệu suất của tiến trình khai phá. Kiểm tra độ chính xác trong việc dự đoán với tập dữ liệu testing.
* **Kết quả tiến trình phân lớp cây quyết định:**
  + Bảng kết quả dự đoán tập dữ liệu sau khi thực hiện cây quyết định :

A screenshot of a computer

Description automatically generated with medium confidence

* Bảng hiệu suất (performance) sẽ thể hiện độ chính xác của việc testing trên tập dữ liệu training (tương đương với số đối tượng dự đoán đúng trên tổng số đối tượng testing). Ngoài ra nó còn cho biết tỷ lệ chính xác khi dự đoán trên từng nhãn.

Table, calendar

Description automatically generated

* Một nhánh nhỏ kết quả cây quyết định Rapidminer biểu diễn:

Diagram

Description automatically generated

### Áp dụng mô hình

* Giới thiệu dữ liệu dùng áp dụng mô hình để dự đoán :
* Lựa chọn dữ liệu: Lấy dữ liệu dataset tại link: <https://archive.ics.uci.edu/ml/datasets/Flags>
* Giới thiệu dữ liệu:
* Tệp dữ liệu này chứa thông tin chi tiết về các quốc gia khác nhau và cờ của họ. Trong tệp này, các trường được phân tách bằng dấu cách (không phải dấu phẩy).
* Dữ liệu gồm 10 thuộc tính có giá trị số. Phần còn lại là Boolean- hoặc giá trị danh nghĩa.
* Thông tin các thuộc tính:

1. Tên: Tên quốc gia liên quan  
   2. Diện tích đất: 1 = N.America, 2 = S.America, 3 = Châu Âu, 4 = Châu Phi, 4 = Châu Á, 6 = Châu Đại Dương  
   3. Khu vực: Góc phần tư địa lý, dựa trên Greenwich và Xích đạo; 1 = NE, 2 = SE, 3 = SW, 4 = NW  
   4. Diện tích: tính bằng nghìn km vuông  
   5. Dân số: tính bằng triệu  
   6. Ngôn ngữ: 1 = tiếng Anh, 2 = tiếng Tây Ban Nha, 3 = tiếng Pháp, 4 = tiếng Đức , 5 = Slavic, 6 = Ấn-Âu khác, 7 = Trung Quốc, 8 = Ả Rập, 9 = Nhật / Thổ Nhĩ Kỳ / Phần Lan / Magyar, 10 = Khác  
   7. tôn giáo: 0 = Công giáo, 1 = Cơ đốc giáo khác, 2 = Hồi giáo, 3 = Phật giáo, 4 = Hindu, 5 = Dân tộc, 6 = Người theo chủ nghĩa Mác, 7 = Người khác  
   8. Thanh: Số thanh dọc trên lá cờ  
   9. Sọc: Số sọc ngang trên lá cờ  
   10. Màu sắc: Số màu khác nhau trong lá cờ  
   11. Đỏ: 0 nếu không có màu đỏ, 1 nếu màu đỏ xuất hiện trên lá cờ  
   12. Xanh lá cây: giống với màu xanh lá cây  
   13. Xanh lam: giống với màu xanh lam  
   14. Vàng: giống với vàng (cũng có màu vàng)  
   15. Trắng: giống với màu trắng  
   16. Màu đen: giống màu đen  
   17. Màu cam: giống màu da cam (cũng có màu nâu)  
   18. Màu mainhue: màu chủ đạo trong cờ (hòa-giải quyết định bằng cách lấy màu trên cùng, nếu không đạt thì đó là màu trung tâm nhất, và nếu không đạt thì sắc độ ngoài cùng bên trái)  
   19. Vòng tròn: Số vòng tròn trên lá cờ  
   20. Dấu chéo: Số đường chéo (thẳng đứng)  
   21. Độ đậm của đường chéo: Số đường chéo  
   22. phần tư: Số phần chia nhỏ  
   23. Biểu tượng: Số biểu tượng mặt trời hoặc ngôi sao  
   24. Lưỡi liềm: 1 nếu có biểu tượng trăng lưỡi liềm, khác 0  
   25. Hình tam giác: 1 nếu có hình tam giác, 0 nếu không  
   26. Biểu tượng: 1 nếu có hình ảnh vô tri vô giác (ví dụ: một chiếc thuyền), nếu không thì 0  
   27. Hình ảnh động: 1 nếu có hình ảnh động (ví dụ: đại bàng, cây cối, bàn tay người), 0 nếu không  
   28. Văn bản: 1 nếu có bất kỳ chữ cái hoặc chữ viết nào trên lá cờ (ví dụ: khẩu hiệu hoặc khẩu hiệu), 0 nếu không thì  
   29. Màu góc trên bên trái: Màu ở góc trên bên trái (di chuyển sang phải để quyết định hòa)  
   30. Màu góc dưới bên trái: Tô màu ở góc dưới bên trái (di chuyển sang trái để quyết định hòa)

* Dữ liệu dùng để áp dụng đã loại bỏ nhãn:

A picture containing text, white

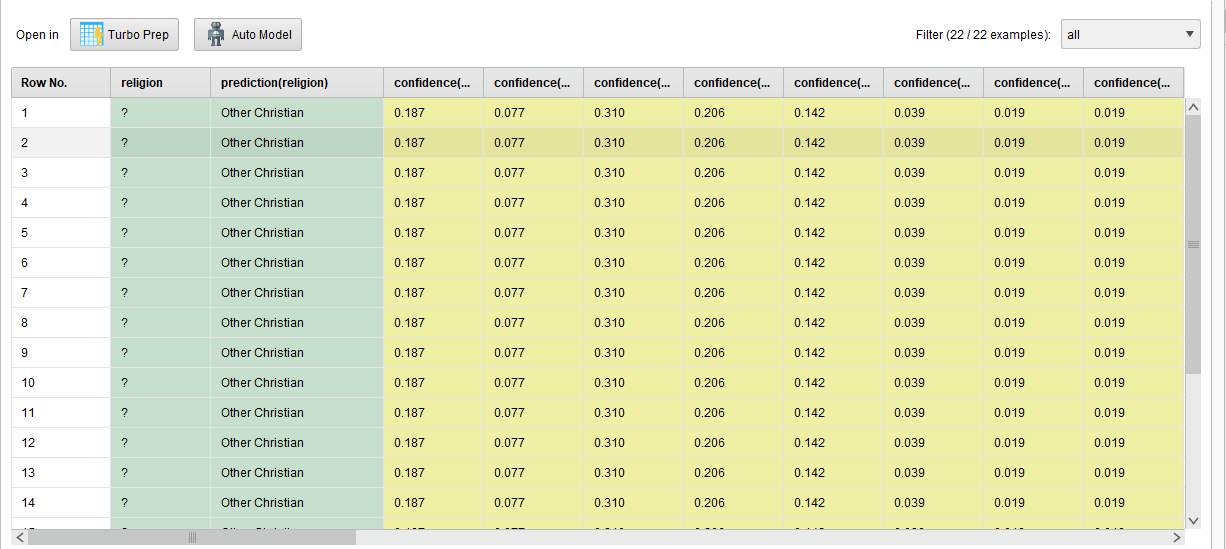
Description automatically generated

* Mô hình áp dụng trên Rapid Miner:

Diagram

Description automatically generated

* Bảng kết quả dự đoán cho tập dữ liệu không có nhãn cũng tương tự như bảng dữ liệu kết quả khi thực hiện training và testing.



* Read excel: Công cụ tử giúp đọc một tập hợp ví dụ từ các tệp bảng tính Excel.
* Numerical to polynominal: Toán tử Numerical to Polynominal được sử dụng để thay đổi kiểu thuộc tính số thành kiểu polynominal, sau đó nó ánh xạ tất cả các giá trị số đơn giản đến các giá trị danh nghĩa tương ứng.
* Set Role: Toán tử này được sử dụng để thay đổi vai trò của một hoặc nhiều thuộc tính.Trong bài toán này chọn thuộc tính Religion để dự đoán.
* Apply model : công cụ giúp áp dụng mô hình thuật toán cây quyết định với dữ liệu testing để dự đoán kết quả của nhãn.
* Bảng kết quả dự đoán cho tập dữ liệu không có nhãn cũng tương tự như bảng dữ liệu kết quả khi testing. Bảng dữ liệu kết quả (example) sẽ bao gồm tất cả những thông tin có trong bảng dữ liệu đầu vào, đồng thời có thêm cột dự đoán (prediction) là kết quả của việc phân lớp cùng với những cột chứa độ tin cậy (confidence) của từng nhãn.

### Nhận xét

* Mô hình cây quyết định dựa vào tập dữ liệu training để xây dựng mô hình cây theo tham số tính toán chính, với thuật toán ID3 là Information Gain còn với thuật toán C4.5 là Gain Ratio.
* Đầu ra của mô hình cây quyết định ngoài bảng dữ liệu kết quả còn có mô hình cây quyết định dựa trên tập dữ liệu training. Qua đó người sử dụng có thể dễ dàng trình bày và giải thích hơn cho những người ít biết về công nghệ.
* Đầu ra của công cụ tính hiệu suất (Performance) là chỉ số chính xác (accuracy) của việc dự đoán với tập dữ liệu testing. Mô hình cây quyết định dự đoán đúng nhãn cho 60.75% tổng số bản ghi trong tập dữ liệu testing.
* Với độ chính xác là 60.75% thì mô hình cây quyết định ở bài toán này hoạt động hiệu quả và hoàn toàn có thể áp dụng để dự đoán.

## ***K- NN***

### Mô tả dữ liệu

* Lựa chọn dữ liệu: Dữ liệu lấy từ đường link <https://archive.ics.uci.edu/ml/datasets/Wine+Quality>
* Giới thiệu dữ liệu: Những dữ liệu này là kết quả phân tích hóa học của rượu vang được trồng trong cùng một vùng ở Ý nhưng có nguồn gốc từ ba giống cây trồng khác nhau. Dữ liệu ghi lại chỉ số của 13 thành phần được tìm thấy trong ba loại rượu vang.
* Các thuộc tính:

1. Alcohol : Loại rượu  
2. Malic acid: Tính axit trong rượu vang  
3. Ash: Hàm lượng tro trong rượu vang.  
4.  Alcalinity of ash: Hàm lượng độ kiềm của tro trong rượu vang.  
5. Magnesium: Một loại hoạt chất cần thiết cho cơ thế chứa trong rượu vang, khoảng 3% Magie.  
6. Total phenols : Tộng hàm lượng phenol trong rượu vang.  
7. Flavanoids : 1 loại chất chống oxy hóa, chất này có trong rượu vang.  
8. Phenol nonflavanoid  
9. Proanthocyanins : Chất chống oxy hóa.  
10. Color intensity: Màu sắc của rượu vang.  
11. Hue   
12. OD280/OD315 of diluted wines: OD280 / OD315 của rượu vang pha loãng  
13. Proline : 1 loại axit amin có trong rượu vang.

* Mục đích: Label dự đoán là Acohol. Dựa vào các chỉ số thuộc tính sẽ dự đoán loại rượu.

### Trình bày ý tưởng thuật toán

* Thuật toán KNN (K – nearest neighbor) là một trong những phương pháp học có giám sát “Superviseed Learning”, hiểu đơn giản là thuật toán sẽ dựa trên biến mục tiêu đã xác định trước đó để xem xét và tìm ra những thuộc tính có thể tác động đến mục tiêu. Thuật toán KNN dựa trên giả định là những thứ tương tự hay có tính chất gần giống nhau sẽ nằm ở vị trí gần nhau nên KNN được xây dựng trên các công thức toán học để tính khoảng cách giữa 2 điểm dữ liệu để xem xét mức độ giống nhau của chúng.
* Tư tưởng chính của thuật toán KNN là tìm cách gán nhãn cho các đối tượng (objects) đã cho dựa theo tập đối tượng mẫu đã gán nhãn có sẵn. Với mỗi đối tượng cần gán nhãn sẽ tìm ra K (K là số được xác đinh trước, K nguyên dương) đối tượng mẫu có khoảng cách gần nhất đến đối tượng. Từ nhãn K đối tượng mẫu lấy ra nhãn xuất hiện nhiều nhất => nhãn của đối tượng.
* Các bước thực hiện:
  + Bước 1: Xác định số K (K là số nguyên dương).
  + Bước 2: Tính khoảng cách giữa đối tượng mới với đối tượng trong tập dữ liệu mẫu (thường dùng khoảng cách Euclidean).
  + Bước 3: Sắp xếp lại khoảng cách nhỏ tới lớn.
  + Bước 4: Lấy K đối tượng khoảng cách nhỏ nhất .
  + Bước 5: Xác định nhãn của đối tượng cần tìm dựa trên nhãn xuất hiện nhiều nhất trong K đối tượng đã chọn.
* Công thức tính khoảng cách hay được sử dụng Euclidean distance: Trong không gian 2 chiều:

A picture containing text, clock, watch

Description automatically generated

* + Trong không gian n chiều: x(x1 ,x2 , …,xn ) và y(y1 ,y2 , …,yn )

Text

Description automatically generated

* + Các thuộc tính cần chuẩn hóa về đơn vị thước đo: Một số cách chuẩn hóa giá trị thuộc tính:
    - Chuyển đổi các giá trị thành các giá trị [0,1]
      * Chia tất cả các giá trị cho giá trị lớn nhất
    - Chuyển đổi Z (Z-transformation):
      * Trừ tất cả các giá trị cho giá trị trung bình, chia cho độ lệch chuẩn .
      * Kết quả được các giá trị [0,1].

### Triển khai mô hình

* Tệp dữ liệu cho dự đoán K – NN: Dữ liệu đầu vào là file excel để training và testing bao gồm 13 thuộc tính chính.

A picture containing calendar

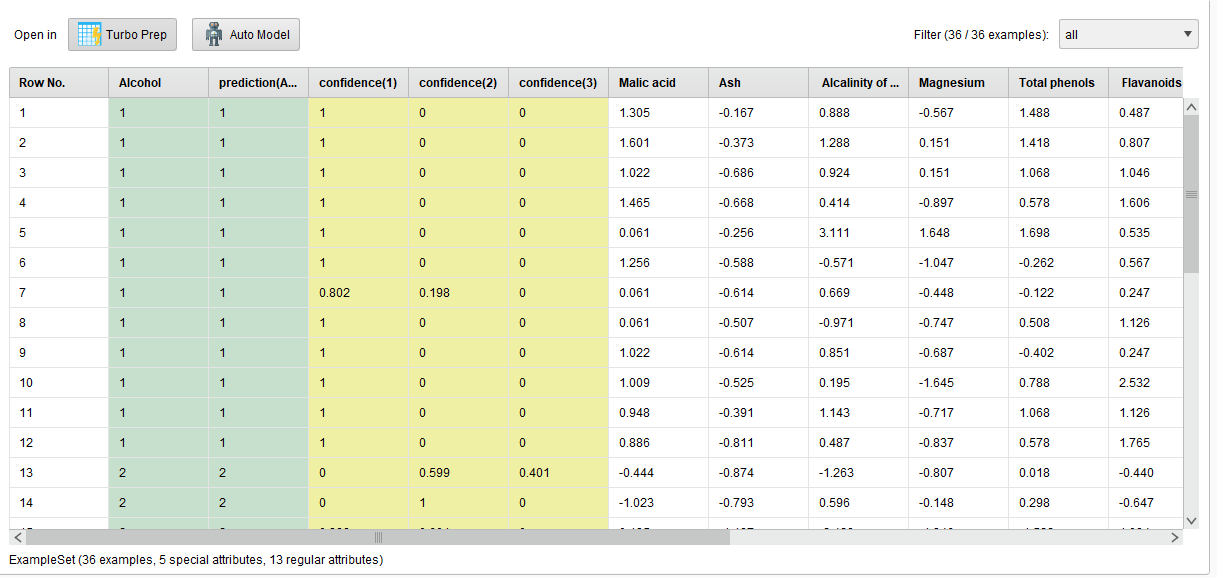
Description automatically generated

* Process triển khai mô hình:

Diagram

Description automatically generated

* Read Execl: Công cụ cho phép đọc file excel có đuôi .xls hoặc .xlxs để đưa tập dữ liệu vào Rapidminer để xử lý.
* Numerical to Polynominal: Toán tử này thay đổi kiểu thuộc tính số đã chọn thành kiểu đa danh. Nó cũng ánh xạ tất cả các giá trị của các thuộc tính này thành các giá trị đa danh nghĩa tương ứng.
* Set role: Công cụ xác định vai trò của các thuộc tính đặc biệt trong thuật toán KNN. Ở đây thuộc tính cần xác định là Alcohol.
* Normalize: Toán tử này chuẩn hóa các giá trị của các thuộc tính đã chọn.Bài toán này chọn phương pháp Z – transformation: Phương pháp chuẩn hóa bằng cách trừ đi trung bình của dữ liệu từ tất cả các giá trị và sau đó phân chia chúng theo độ lệch chuẩn. Sau đó, phân phối dữ liệu có trung bình 0 và phương sai 1.
* Split data: Công cụ tách tập dữ liệu gốc ra thành 2 phần với tỷ lệ mong muốn. Một phần sẽ thực hiện việc training và phần còn lại sẽ thực hiện việc testing. Trong bài toán này tạo ra 2 tập con : 80% dữ liệu cho training và 20% dữ liệu cho testing.
* K – NN: Công cụ thực hiện mô hình thực hiện thuật toán KNN với tập dữ liệu training. Chọn chỉ số láng giềng gần nhất ở thuật toán này giá trị K = 5.
* Apply Model: Công cụ giúp áp dụng mô hình thuật toán k-NN với dữ liệu testing để dự đoán kết quả của Alcohol (label).
* Performance: Công cụ thực hiện tính hiệu suất của tiến trình khai phá. Kiểm tra độ chính xác trong việc dự đoán với tập dữ liệu testing.
* **Kết quả tiến trình phân lớp sử dụng thuật toán K - NN:**
* Bảng dữ liệu kết quả (example) sẽ bao gồm tất cả những thông tin có trong bảng dữ liệu đầu vào, và có thêm cột dự đoán (prediction) là kết quả của việc phân lớp cùng với những cột chứa độ tin cậy (confidence) của từng nhãn khi thực hiện việc phân lớp. Rapidminer sẽ chọn nhãn có độ tin cậy cao nhất là kết quả dự đoán.



* Bảng hiệu suất (performance) sẽ thể hiện độ chính xác của việc testing trên tập dữ liệu training (tương đương với số đối tượng dự đoán đúng trên tổng số đối tượng testing). Ngoài ra nó còn cho biết tỷ lệ chính xác khi dự đoán trên từng nhãn.

Graphical user interface

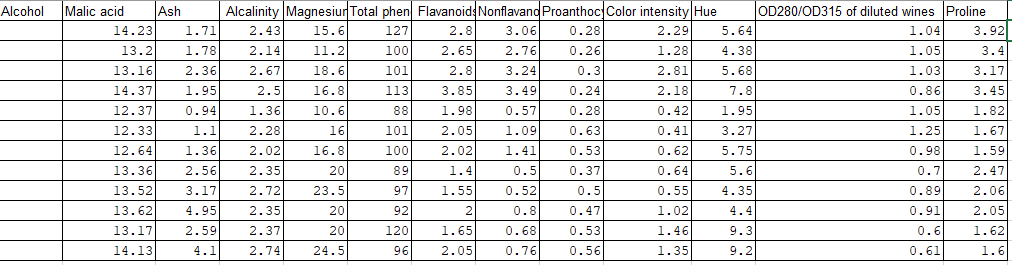
Description automatically generated with low confidence

### Áp dụng mô hình:

* Giới thiệu dữ liệu: Những dữ liệu này là kết quả phân tích hóa học của rượu vang được trồng trong cùng một vùng ở Ý nhưng có nguồn gốc từ ba giống cây trồng khác nhau. Dữ liệu ghi lại chỉ số của 13 thành phần được tìm thấy trong ba loại rượu vang.
* Các thuộc tính:

1. Alcohol : Loại rượu  
2. Malic acid: Tính axit trong rượu vang  
3. Ash: Hàm lượng tro trong rượu vang.  
4.  Alcalinity of ash: Hàm lượng độ kiềm của tro trong rượu vang.  
5. Magnesium: Một loại hoạt chất cần thiết cho cơ thế chứa trong rượu vang, khoảng 3% Magie.  
6. Total phenols : Tộng hàm lượng phenol trong rượu vang.  
7. Flavanoids : 1 loại chất chống oxy hóa, chất này có trong rượu vang.  
8. Phenol nonflavanoid  
9. Proanthocyanins : Chất chống oxy hóa.  
10. Color intensity: Màu sắc của rượu vang.  
11. Hue   
12. OD280/OD315 of diluted wines: OD280 / OD315 của rượu vang pha loãng  
13. Proline : 1 loại axit amin có trong rượu vang.

* Data áp dụng để triển khai dự đoán đã được loại bỏ nhãn:

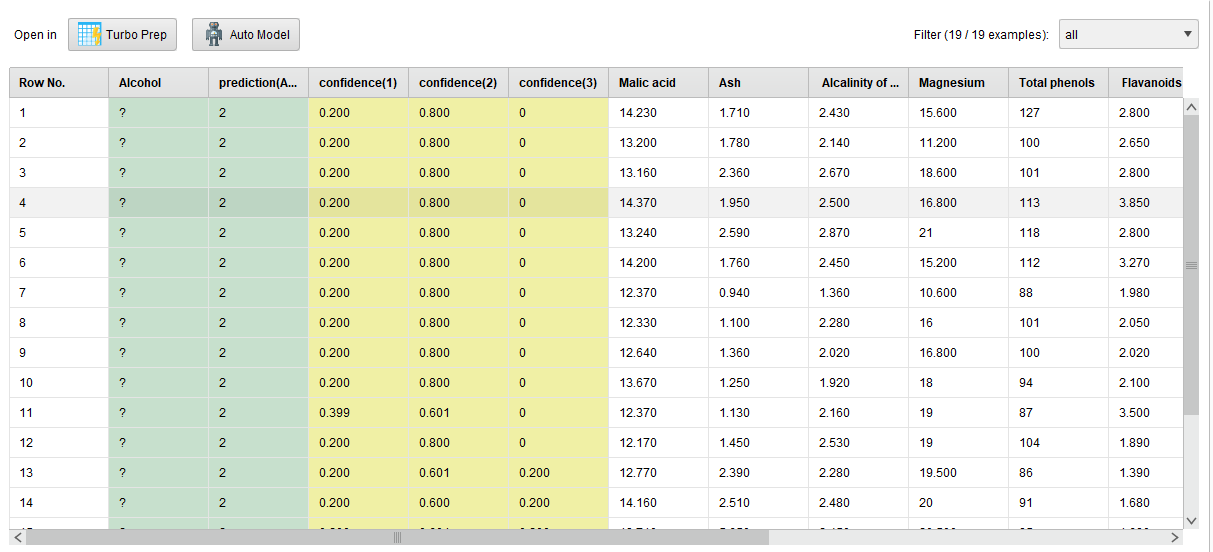


* Mô hình áp dụng:

Diagram, schematic

Description automatically generated

* Bảng kết quả dự đoán cho tập dữ liệu không có nhãn cũng tương tự như bảng dữ liệu kết quả khi thực hiện training và testing.



### Nhận xét

* Mô hình k-NN cho phép xử lý cả dữ liệu kiểu chữ và dữ liệu kiểu số. Với dữ liệu kiểu chữ thì khoảng cách sẽ là 0 (khác nhau) hoặc 1 (giống nhau).
* Đầu ra của công cụ tính hiệu suất Performance là chỉ số chính xác của việc dự đoán tập dữ liệu testing. Trong bài toán này độ chính xác lên đến 97.22%.
* Với độ chính xác 97.22%, chúng ta có thể xử dụng thuật toán K- NN để dữ đoán cho tập dữ liệu chưa có nhãn.