MỤC LỤC

[I.Thông tin chung 3](#_Toc130686985)

[1.Thông tin thành viên: 3](#_Toc130686986)

[2.Thông tin giảng viên: 3](#_Toc130686987)

[3.Môi trường và ngôn ngữ làm việc: 3](#_Toc130686988)  
[4.Độ hoàn thành: 3](#_Toc130686989)

[II.Nội dung 4](#_Toc130686990)

[1.Yêu cầu 1: Install Weka: 4](#_Toc130686991)

[1.1.Giải thích ý nghĩa các nhóm điều khiển trong tab PreProcess: 4](#_Toc130686992)

[1.2.Giải thích ngắn gọn ý nghĩa 5 tab trong giao diện Explorer của Weka: 4](#_Toc130686993)

[2.Yêu cầu 2: Getting Acquainted With WEKA: 5](#_Toc130686994)

[2.1.Exploring Breast Cancer data set: 5](#_Toc130686995)

[2.2.Exploring Weather data set: 9](#_Toc130686996)

[2.3.Exploring Credit in Germany data set: 13](#_Toc130686997)

[3.Yêu cầu 3: 21](#_Toc130686998)

1. **Thông tin chung**
2. **Thông tin thành viên:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | MSSV | Họ và tên | Email | Tỉ lệ đóng góp |
| 1 | 20127186 | Nguyễn Hoàng Gia Huy | 20127186@student.hcmus.edu.vn | 100% |
| 2 | 20127586 | Nguyễn Đình Pháp | [20127586@student.hcmus.edu.vn](mailto:20127586@student.hcmus.edu.vn) | 100% |

1. **Thông tin giảng viên:**

* Thầy Lê Hoài Bắc.
* Thầy Nguyễn Bảo Long.
* Cô Nguyễn Thị Thu Hằng.

1. **Môi trường và ngôn ngữ làm việc:**

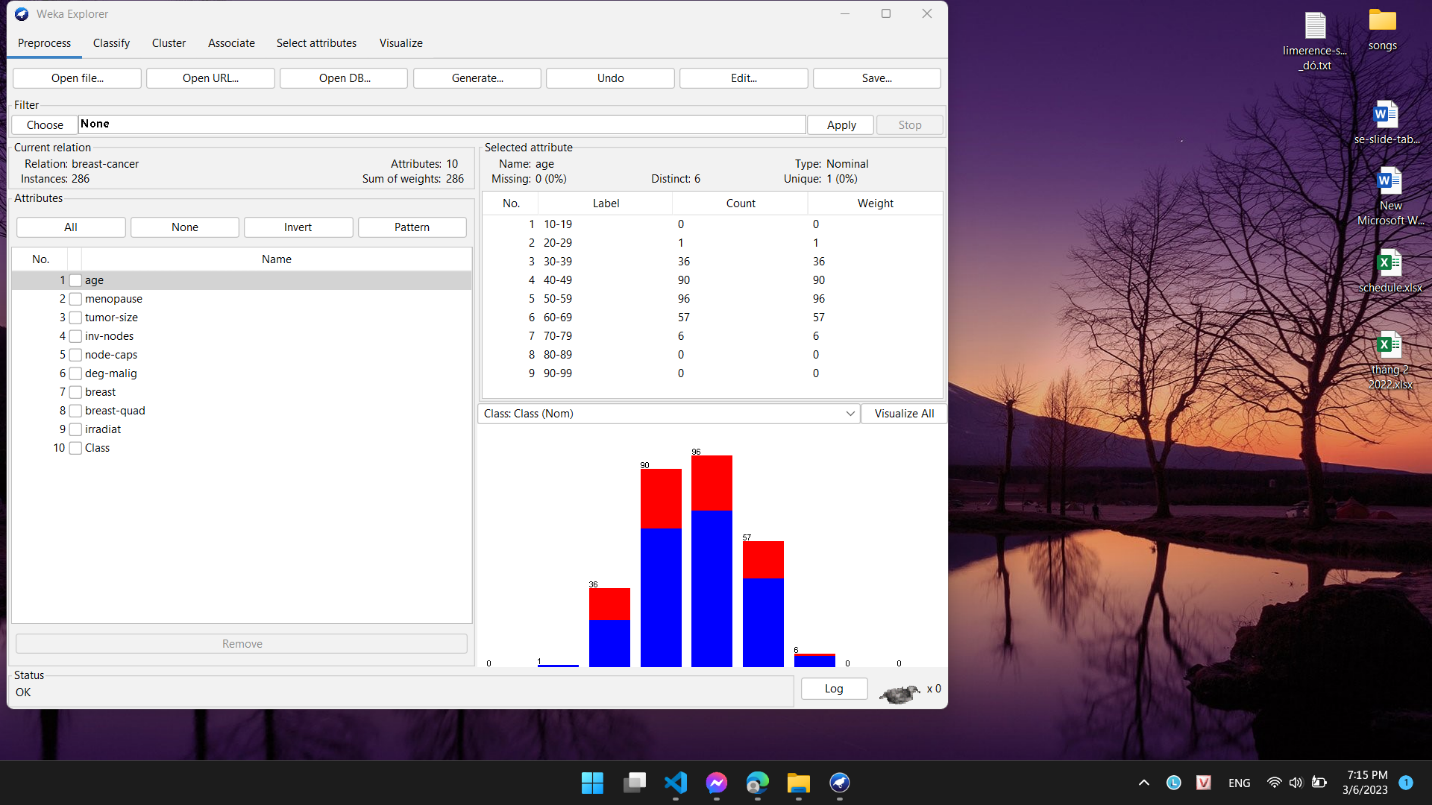
* Visual studio code (version: 1.76.0)
* Weka (version: weka-3-8-6-azul-zulu-windows)
* Python 3

1. **Độ hoàn thành:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | MSSV | Yêu cầu | | Mức độ hoàn thành |
| 1 | 20127186 20127586 | Yêu cầu 1 | Install WEKA | 100% |
| 2 | 20127586 | Yêu cầu 2:Getting Acquainted With WEKA | 2.1: Exploring Breast Cancer data set | 100% |
| 3 | 20127586 | 2.2: Exploring Weather data set | 100% |
| 4 | 20127186 | 2.3: Exploring Credit in Germany data set | 100% |
| 5 | 20127586 | Yêu cầu 3: Preprocessing Data in Python | 3.1: Extract columns with missing values | 100% |
| 6 | 20127586 | 3.2: Count the number of lines with missing data | 100% |
| 7 | 20127586 | 3.3: Fill in the missing value using mean, median (for numeric properties) and mode (for the categorical attribute) | 100% |
| 8 | 20127186 | 3.4: Deleting rows containing more than a particular number of missing values | 100% |
| 9 | 20127186 | 3.5: Deleting columns containing more than a particular number of missing values | 100% |
| 10 | 20127186 | 3.6: Delete duplicate samples | 100% |
| 11 | 20127586 | 3.7: Normalize a numeric attribute using min-max and Z-score methods | 100% |
| 12 | 20127186 | 3.8: Performing addition, subtraction, multiplication, and division between two numerical attributes | 100% |

1. **Nội dung**
2. **Yêu cầu 1: Install Weka:**

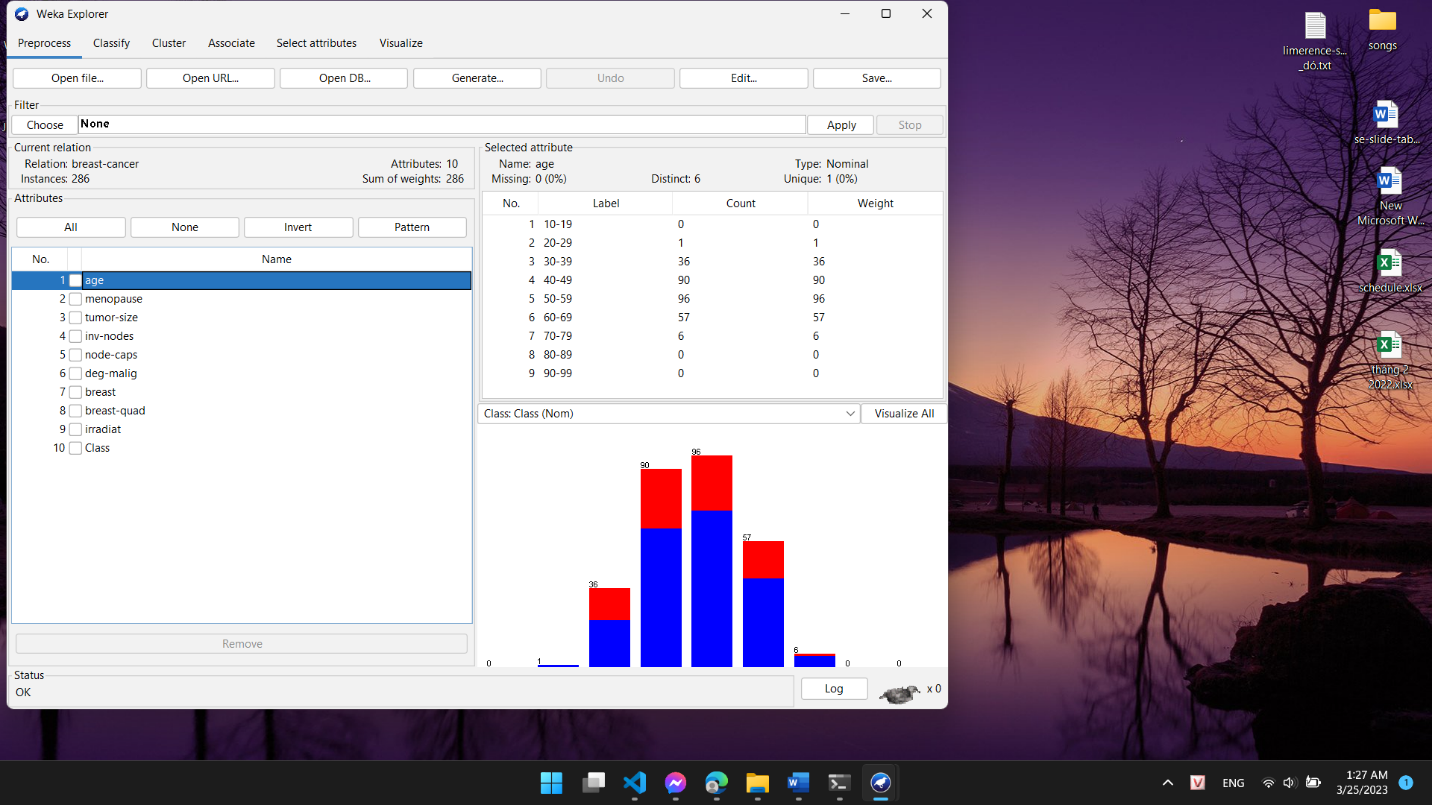
Hình ảnh giao diện chức năng Explorer cùng màn hình desktop cá nhân:



*Hình ảnh*: Giao diện chức năng Explorer của Weka

* 1. **Giải thích ý nghĩa các nhóm điều khiển trong tab PreProcess:**
* Current relation: Cho biết các thông tin chung về tập dữ liệu như tên của tập dữ liệu, số mẫu, số thuộc tính.
* Attributes: Cho biết danh sách các thuộc tính hiện tại có trong tập dữ liệu.
* Selected attribute: Cho biết thông tin liên quan đến thống kê của thuộc tính đang được chọn ở phần Attributes, như trung bình cộng, gía trị min – max. Đối với dạng dữ liệu định danh, Weka sẽ cung cấp danh sách định danh và số lượng mỗi định danh.
  1. **Giải thích ngắn gọn ý nghĩa 5 tab trong giao diện Explorer của Weka:**
* PreProcess: Tiền xử lý dữ liệu.
* Classify: Phân lớp xử lý.
* Cluster: Gom cụm dữ liệu.
* Associate: Khai phá các luật kết hợp.
* Select attributes: Lựa chọn thuộc tính. Sử dụng khi cần xem xét mối tương quan giữa các thuộc tính.
* Visualize: Trực quan hóa dữ liệu.

1. **Yêu cầu 2: Getting Acquainted With WEKA:**
   1. **Exploring Breast Cancer data set:**



*Hình ảnh*: Tập dữ liệu breast-cancer trên giao diện Explorer của Weka

Trả lời cho các câu hỏi:

* How many instances does this data set have? (Tập dữ liệu có bao nhiêu mẫu?)

Application

Description automatically generated with low confidence

* Có 286 instances.
* How many attributes does this data set have? (Tập dữ liệu có bao nhiêu thuộc tính?)

Application

Description automatically generated with low confidence

* Có 10 attributes.
* Which attribute is used for the label? Can it be changed? How? (Thuộc tính nào được dùng làm lớp – class? Có thể thay đổi thuộc tính dùng lớp hay không? Nếu được thì bằng cách nào?)
* Thuộc tính được dùng làm class là thuộc tính “class” có 2 giá trị recurrence-events và no-recurrence-events. Có thể thay đổi thuộc tính lớp (class). Cách thực hiện: chọn Edit, nhấp chuột phải tại thuộc tính muốn làm class, chọn Attribute as class.

Graphical user interface

Description automatically generated



Graphical user interface, table

Description automatically generated

Và nhấn chọn “OK”.

* What is the meaning of each attribute? (Ý nghĩa từng thuộc tính là gì?)
* Có 10 thuộc tính:
  + Age: khoảng độ tuổi.
  + Menopause: thuộc tính thể hiện giai đoạn của mãn kinh.
    - Premeno: giai đoạn tiền mãn kinh.
    - Ge40: viết tắt của “grade and ER/PR status under 40”, một chỉ số cho thấy mức độ nghiêm trọng của ung thư vú.
    - Lt40: phụ nữ dưới 40.
  + Tumor-size: thuộc tính thể hiện kích thước của khối u.
  + Inv-nodes: thuộc tính thể hiện kích thước của các núm vú bị xâm lấn.
  + Node-caps: thuộc tính thể hiện có hay không việc có màng bọc núm vú.
  + Deg-malig: viết tắt của "Degree of malignancy" (độ ác tính). Thuộc tính này thể hiện mức độ nghiêm trọng của tế bào ung thư, dựa trên đánh giá về mức độ khác biệt về hình dạng và cấu trúc giữa tế bào ung thư và các tế bào bình thường trong mẫu mô bị tác động.
  + Breast: thuộc tính thể hiện bệnh nhân ung thư vú bên trái hay phải.
  + Breast-quad: thuộc tính thể hiện vùng ngực bị ung thư vú của bệnh nhân.
  + Irradiat: thuộc tính thể hiện bệnh nhân có xử dụng xạ trị hay không.
  + Class: thuộc tính thể hiện kết quả của bệnh nhân sau khi điều trị thành công là recurrence-events (tái phát) và no-recurrence-events (không tái phát).
* Let’s investigate the missing value status in each attribute and describe in general ways to solve the problem of missing values. (Hãy điều tra những “missing value” trong mỗi thuộc tính và trình bày tổng quát các cách giải quyết của việc “missing value”.)
* Có 2 thuộc tính gặp trường hợp “missing value” là node-caps và breast-quad.

Table, timeline

Description automatically generated with medium confidence

Table

Description automatically generated

Trong đó, thuộc tính node-caps bị missing value nhiều nhất (là 8 mẫu chiếm 3%), và thuộc tính breast-quad bị missing value ít nhất (là 1 mẫu chiếm gần như là 0%).

Giải pháp tổng quát cho việc “missing value” như sau:

* + Có thể loại bỏ dữ liệu thiếu đó hoặc bổ sung dữ liệu thiếu đó bằng các phương pháp như k-nearest neighbors (KNN),…
* Let’s propose solutions to the problem of missing values in the specific attribute. (Hãy đề xuất các giải pháp cho trường hợp missing value trong thuộc tính cụ thể)
* Giải pháp:
  + Đối với thuộc tính breast\_quad: vì chỉ “missing value” 1 đơn vị 🡪 không ảnh hưởng nhiều tới dữ liệu 🡪 xóa dữ liệu này.
  + Đối với thuộc tính node-caps: vì ảnh hưởng đến 3% nên có thể chọn giải pháp sử dụng mô hình dự đoán giá trị thiếu dựa trên các thuộc tính khác của dữ liệu.
* Let’s explain the meaning of the chart in the WEKA Explorer. Setting the title for it and describing its legend. (Hãy giải thích ý nghĩa của biểu đồ trong Weka Explorer. Đặt tiêu đề cho nó và mô tả nó.)
* Ý nghĩa của biểu đồ:
  + Màu xanh biểu thị tại mỗi khoảng dữ liệu của thuộc tính được chọn, có bao nhiêu mẫu cho kết quả no-recurrence-events, và tương tự màu đỏ cho recurrent-events.
  + Đặt tiêu đề: Đồ thị phân lớp.

Chart, bar chart

Description automatically generated

* 1. **Exploring Weather data set**:

Graphical user interface, application

Description automatically generated

*Hình ảnh:* Tập dữ liệu breast-cancer trên giao diện Explorer của Weka

Trả lời cho các câu hỏi:

* How many attributes does this data set have? How many samples? Which attributes have data type categorical? Which attributes have a data type that is numerical? Which attribute is used for the label? (Tập dữ liệu này có bao nhiêu thuộc tính? Có bao nhiêu mẫu? Thuộc tính nào có kiểu dữ liệu phân loại? Thuộc tính nào có kiểu dữ liệu là số? Thuộc tính nào được sử dụng cho nhãn?)
* Trả lời:

Graphical user interface

Description automatically generated with low confidence

Graphical user interface, table

Description automatically generated

* + Tập dữ liệu có 5 thuộc tính và 14 mẫu.
  + Thuộc tính có kiểu dữ liệu phân loại (categorical): Không có.
  + Thuộc tính có kiểu dữ liệu số (numerical): temperature và humidity.
  + Thuộc tính sử dụng cho nhãn (label) là “play”.

Bar chart

Description automatically generated

* Let’s list ***five-number summary*** of two attributes ***temperature*** and ***humidity***. Does WEKA provide these values? (Hãy liệt kê 5 số tóm tắt của hai thuộc tính nhiệt độ và độ ẩm. WEKA có cung cấp những giá trị này không?)
* Five-number summary of two attributes temperature and humidity.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attribute | Min | Q1 | Median | Q3 | Max |
| Temperature | 64 | 69 | 72 | 80 | 85 |
| Humidity | 65 | 70 | 82.5 | 90 | 96 |

Graphical user interface, application

Description automatically generated

Graphical user interface

Description automatically generated

* Let’s explain the meaning of all charts in the WEKA Explorer. Setting the title for it and describing its legend. (Hãy giải thích ý nghĩa của tất cả các biểu đồ trong WEKA Explorer. Đặt tiêu đề cho nó và mô tả nó)

Chart, bar chart

Description automatically generated

Đồ thị biểu diễn thuộc tính của tập dữ liệu weather

* Let’s move to the Visualize tag. What’s the name of this chart? Do you think there are any pairs of different attributes that have correlated? (Hãy chuyển sang thẻ Visualize. Tên của biểu đồ này là gì? Bạn có nghĩ rằng có bất kỳ cặp thuộc tính khác nhau có tương quan?)
* Trả lời:

A picture containing calendar

Description automatically generated

Đồ thị phân tán của các thuộc tính trong tập dữ liệu Weather

Tên biểu đồ là biểu đồ scatter (scatter plot).

Dựa vào các đồ thị quan sát được, các trường dữ liệu chỉ tương quan với chính bản thân. Còn lại các trường khác với nhau không có tương quan.

* 1. **Exploring Credit in Germany data set:**

Graphical user interface

Description automatically generated with medium confidence

Trả lời cho các câu hỏi sau:

* What is the content of the comments section in credit-g.arff (when opened with any text editor) about? How many samples does the data set have? How many attributes? Describe any five attributes (must have both discrete and continuous attributes). (Nội dung của phần nhận xét trong credit-g.arff (khi được mở bằng bất kỳ trình soạn thảo văn bản nào) nói về điều gì? Tập dữ liệu có bao nhiêu mẫu? Có bao nhiêu thuộc tính? Mô tả năm thuộc tính bất kỳ (phải có cả thuộc tính rời rạc và liên tục))
* Trả lời:

Text

Description automatically generated

* + Khi được mở bằng bất kỳ text editor nào sẽ có mô tả gồm: tiêu đề (title), thông tin về nguồn gốc của tập dữ liệu, số lượng mẫu, thuộc tính,…
  + Tập dữ liệu này có 1000 mẫu (instances) và 21 thuộc tính (attributes).

Graphical user interface

Description automatically generated with medium confidence

* + 5 thuộc tính bất kỳ:
    - duration (thuộc tính liên tục): thời hạn vay tín dụng, đơn vị tháng.
    - credit\_history (thuộc tính rời rạc): lịch sử giao dịch tín dụng, có 5 giá trị: critical/other existing credit, existing paid, delayed previously, no credits/all paid, all paid.
    - purpose (thuộc tính rời rạc): mục đích của các khoản vay tín dụng, có 11 giá trị:
      * New car
      * Used car
      * Furniture/equipment
      * Radio/tv
      * Domestic appliance
      * Repairs
      * Education
      * Vacation
      * Retraining
      * Business
      * Other
    - saving\_status (thuộc tính liên tục): tình trạng tiết kiệm của đơn vị dữ liệu đó, gồm các mức giá trị sau:
      * Nhỏ hơn 100 tiền Đức
      * Trong khoảng 100 đến 500 tiền Đức
      * Trong khoảng 500 đến 1000 tiền Đức
      * Lớn hơn 1000 tiền Đức
    - employment (thuộc tính rời rạc): tình trạng việc làm của khách hàng, gồm các giá trị định danh như
      * unemployed
      * <= 1 year
      * 1<= employment <= 4 years
      * > 4 years
* Which attribute is used for the label? (thuộc tính nào được dùng làm nhãn?)
* Trả lời:

Thuộc tính “class” được dùng làm label.

* Let’s describe the distribution of continuous attributes? (Left skewed or right skewed?) (Hãy mô tả sự phân bố của thuộc tính liên tục? (Xiên trái hay lệch phải?))
* Trả lời:

Dựa vào biểu đồ của 1 số thuộc tính liên tục như duration, age, credit\_amount, chúng em thấy được sự phân bố lệch về phía bên trái.

Chart

Description automatically generatedChart

Description automatically generatedChart, waterfall chart

Description automatically generated

* Let’s explain the meaning of all charts in the WEKA Explorer. Setting the title for it and describing its legend.
* Trả lời:

Timeline

Description automatically generated

Đồ thị biểu diễn thuộc tính của tập dữ liệu credit\_german

* Let’s move to the Select attributes tag. Describe all of the options for attribute selection.
* Trả lời :

Graphical user interface, text, application, email

Description automatically generated

* + Attribute Evaluator: This option is used to evaluate the individual attributes based on the selected criteria such as the correlation with the class, information gain, or chi-squared test. The attribute evaluator ranks the attributes based on their importance to the classification task.
  + Search Method: This option is used to search for the subset of attributes that provides the best classification accuracy. The search methods available in WEKA include Best First, Greedy Stepwise, and Genetic Search.
  + Evaluation: This option is used to evaluate the performance of the selected attributes. The evaluation method can be cross-validation or percentage split.
  + Start Set: This option is used to specify the set of attributes that should be considered for selection. By default, all attributes in the dataset are considered.
  + Output calculated attribute evaluations: This option is used to output the evaluation scores of each attribute.
  + Output selected attributes: This option is used to output the final set of selected attributes.
  + Attribute selection mode: This option is used to specify the mode of attribute selection. The options available in WEKA include selecting a fixed number of attributes, selecting attributes above a certain threshold score, or selecting the top percentage of attributes.
* Which options should be used to select the 5 attributes with the highest correlation?(Step-by-step description, with step-by-step photos and final results)
* Trả lời:

Chọn GainRatioAttributeEval hay InfoGainAttributeEval đều cho biết về các thuộc tính có tương quan cao nhất đối với thuộc tính lớp. Ở đây, em chọn InfoGainAttributeEval để chọn các thuộc tính này.

* + Bước 1: vào tab **Select attributes.** Ở mục **Attribute Evaluator**, chọn **InfoGain AttributeEval**. Tại mục **Search Method**, chọn **Ranker.**

Graphical user interface, text, application, email

Description automatically generated



* + Bước 2: Bấm start. Kết quả xuất hiện.

Graphical user interface, text, application, email

Description automatically generated



Từ kết quả trên, ta có thể thấy 5 thuộc tính có tương quan cao nhất với thuôc tính class là checking\_status, credit\_history, duration, saving\_status, purpose.

1. **Yêu cầu 3:**

zzzzz