ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN



DATA MINING

LAB 1: PREPROCESSING

*Lớp*: KTDL & UD 18\_21

*Nhóm thực hiện*:

18120078 – Ngô Phù Hữu Đại Sơn

18120253 – Mai Ngọc Tú

MỤC LỤC

[A. Thông tin khái quát 2](#_Toc54517123)

[I. Thông tin nhóm 2](#_Toc54517124)

[II. Bảng phân công công việc 2](#_Toc54517125)

[B. Nội dung 3](#_Toc54517126)

[I. Mục tiêu của đồ án: 3](#_Toc54517127)

[II. Yêu cầu: 3](#_Toc54517128)

[III. Triển khai 4](#_Toc54517129)

[1. Cài đặt Weka: 4](#_Toc54517130)

[2. Làm quen với Weka: 4](#_Toc54517131)

[3. Cài đặt tiền xử lý dữ liệu: 4](#_Toc54517132)

[IV. Đánh giá đồ án 4](#_Toc54517133)

[1. Mức độ hoàn thành của các thành viên 4](#_Toc54517134)

[2. Mức độ hoàn thành đồ án: 4](#_Toc54517135)

[VII. Nguồn tham khảo 4](#_Toc54517136)

# Thông tin khái quát

## I. Thông tin nhóm

|  |  |  |
| --- | --- | --- |
| **MSSV** | **Họ tên** | **Vai trò** |
| 18120078 | Ngô Phù Hữu Đại Sơn | Nhóm trưởng |
| 18120253 | Mai Ngọc Tú | Thành viên |

## II. Bảng phân công công việc

|  |  |  |
| --- | --- | --- |
| **MSSV** | **Công việc phụ trách** | **Thời gian thực hiện** |
|  |  |  |
|  |  |  |

# B. Nội dung

## I. Mục tiêu của đồ án:

* Làm quen với các thao tác cơ bản trong tác vụ tiền xử lý dữ liệu thông qua việc áp dụng các công cụ hỗ trợ được cung cấp bởi phần mềm mã nguồn mở Weka.
* Phát huy kỹ năng lập trình để tự cài đặt các thủ tục tiền xử lý dữ liệu đơn giản.

## II. Yêu cầu:

1. **Cài đặt Weka.**

* Chụp hình giao diện chức năng Exploere cùng màn hình desktop.
* Tìm thư mục data trong thư mục cài đặt của Weka và mở một tập dữ liệu bất kì (có phần mở rộng là arff). Giải thích ý nghĩa các nhóm điều khiển Current relation, Attributes và Selected attribute trong tab Preprocess. Giải thích ngắn gọn ý nghĩa 5 tab trong giao diện Explorer của Weka.

1. **Làm quen với Weka.**

* Khám phá bộ dữ liệu Breast Cancer
* Khám phá bộ dữ liệu Weather
* Khám phá bộ dữ liệu Tín dụng Đức

1. **Cài đặt tiền xử lý dữ liệu**

* Cài đặt chương trình khám phá bộ dữ liệu house\_prices gồm có các chức năng:

+ Liệt kê các cột bị thiếu dữ liệu.

+ Đếm số dòng bị thiếu dữ liệu.

+ Điền giá trị bị thiếu.

+ Xóa các dòng bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước.

+ Xóa các cột bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước.

+ Xóa các mẫu bị trung lặp.

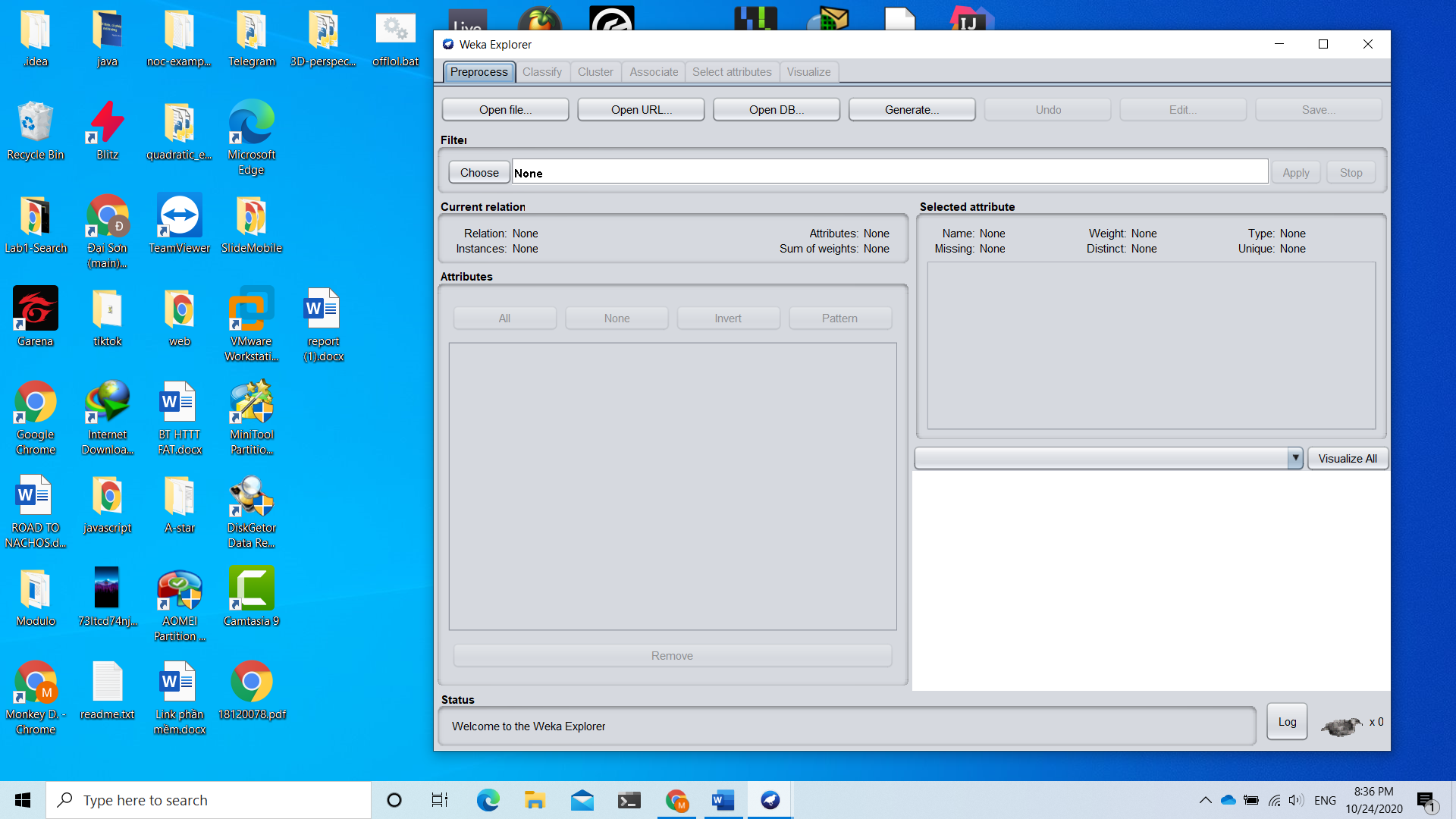
+ Chuẩn hóa một thuộc tính numeric bằng phương pháp min-mã và Z-score.

+ Tính giá trị biểu thức thuộc tính.

## III. Triển khai

### Cài đặt Weka:

* Giao diện chức năng Explorer:



* Giải thích ý nghĩa các nhóm điều khiển
* Current Relation: Cho biết thông tin về tập dữ liệu đang khảo sát

+ Relation: Tên của quan hệ

+ Instances: Số mẫu

+ Attributes: Số thuộc tính

* Attributes: Danh sách các thuộc tính được hiển thị theo đúng thứ tự khai báo ở trong file
* Selected attribute: Liệt kê các giá trị tương ứng với thuộc tính đang được chọn ở nhóm Attributes
* Giải thích ý nghĩa các tab trong giao diện Explorer:
* Preprocessing: có chức năng lọc dữ liệu, mặc khác cung choc ho người dung các thông tin về tập dữ liệu đang xử lý.
* Classify: có chức năng phân lớp.
* Cluster: có chức năng gom cụm.
* Select Attributes: chọn các thuộc tính thích hợp nhất trong dữ liệu.
* Associate: có chức năng rút ra các luật kết hợp.
* Visualize: có chức năng trực quan hóa dữ liệu.

### Làm quen với Weka:

* 1. **Đọc dữ liệu vào Weka:**
  2. **Khám phá tập dữ liệu Weather:**
  3. **Khám phá tập dữ liệu Tín dụng Đức:**

1. *Nội dung của phần ghi chú (Comment) trong credit\_g.arff nói về:*

+ Tên của tập dữ liệu: German credit data

+ Thông tin về tác giả (Professor Dr. Hans Hofmann)

+ Số lượng mẫu khảo sát: 1000

+ Có 2 bộ dữ liệu được cung cấp. Bản gốc do Giáo sư Hofmann cung cấp, chứa các thuộc tính phân loại nằm trong file “german.data”

+ Đối với các thuật toán cần thuộc tính số học, Đại học Strathclyde đã cung cấp tệp "german.data-numeric". Tệp này đã được chỉnh sửa và thêm một số biến chỉ số để phù hợp với các thuật toán không thể đối phó với các biến phân loại. Một số thuộc tính phân loại có thứ tự (chẳng hạn như thuộc tính thứ 17) đã được mã hóa dưới dạng số nguyên.

+ Số thuộc tính của “german”: 21 (7 số học, 13 phân loại và 1 thuộc tính lớp)

+ Số thuộc tính của “german.numer”: 24 (24 số học)

+ Các thuộc tính được mô tải cho người Đức (german)

+ Ma trận chi phí: (các dòng thể hiện phân lớp trong thực tế và cột thể hiện dự đoán phân lớp.

+ Gán lại tên các thuộc tính thành các tên có ý nghĩa

+ Mô tả 5 thuộc tính:

* + - *Thuộc tính 1*: tình trạng của các tài khoản thanh toán hiện tại
      * A11: < 0 DM (Deutsche Mark, 1 đơn vị tiền tệ, quy đổi khoảng 90 cents Canada)
      * A12: 0 <=… < 20 DM
      * A13: … >= 200 DM
      * A14: không có tài khoản thanh toán
    - *Thuộc tính 2*: Thời gian trong tháng
    - *Thuộc tính 3*: Lịch sử tín dụng:
      * A30: không có tín dụng nào được thực hiện hoặc tất các các tín dụng đã được chi trả hợp lệ
      * A31: Tất cả các tín dụng tại ngân hàng này đã được chi trả hợp lệ
      * A32: existing credits paid back duly till now
      * A33: delay in paying off in the past
      * A34: critical account/ other credits existing (not at this bank)
    - *Thuộc tính 4*: Mục đích vay tiền:
      * Mua xe mới
      * Mua xe cũ
      * Mua nội thất
      * Mua TV/Radio
      * Mua Đồ gia dụng
      * Sửa chữa
      * Giáo dục
      * Du lịch
      * Huấn luyện lại
      * Kinh doanh
      * Khoản khác
    - *Thuộc tính 13*: Tuổi

1. *Tên của thuộc tính lớp là gì? Đánh giá phân bố của các lớp, tức là cân bằng hay lệch về một lớp?*

+ Tên thuộc tính lớp là “class”

+ Phân bố lớp không đồng đều. Các đánh giá tốt nhiều hơn các đánh giá không tốt (7:3)

1. *Sử dụng tab Select attributes. Liệt kê những lựa chọn khác nhau của Weka để chọn lọc thuộc tính, giải thích ngắn gọn từng phương pháp*.

*+ Để chọn lọc thuộc tính trong weka, gần xác định Attibute Evaluator và Search Method.*

*+ Attribute Evaluator:*

* + - *Correlation-based Feature Subset Selection(CfsSubsetEval):* đánh giá giá trị của tập con các thuộc tính bằng cách xem xét khả năng dự đoán riêng của từng đối tượng cùng với mức độ dư thừa giữa chúng.
    - *Classifier Atribute Evaluator (ClassifierAttributeEval):* Đánh giá giá trị của tập con các thuộc tính bằng cách sử dụng bộ phân loại do người dùng chỉ định.
    - *Classifier Subset Evaluator (ClassifierSubsetEval):* đánh giá tập con các thuộc tính dựa trên tập huấn luyện.
    - *GainRatioAttributeEval:* Chọn các thuộc tính có độ đo Ratio Gain cao.
    - OneRAtributeEval: đánh giá tập con các thuộc tính dựa trên bảng tần số giữa tập thuộc tính với thuộc tính lớp*.*
    - *Principal components Analysis (PCA):* chọn tập thuộc tính không có tương quan với nhau và các thuộc tính trong tập này là tổ hợp tuyến tính của các thuộc tính ban đầu.
    - *ReliefFAttributeEval:* Đánh giá giá trị của một thuộc tính bằng cách liên tục lấy mẫu một đối tượng và xem xét giá trị của thuộc tính đã cho với đối tượng gần nhất của cùng một lớp và khác lớp.
    - *Correlation Based Feature Selection (CorrelationAttributeEval):* Chọn các thuộc tính có tính tương quan với thuộc tính lớp cao. Kết quả trả về số nguyên nằm trong đoạn [-1;1].

+ Positice Correlation(1): 2 thuộc tính tỉ lệ thuận với nhau

+ Neutral Correlation(0): 2 thuộc tính không tương quan với nhau

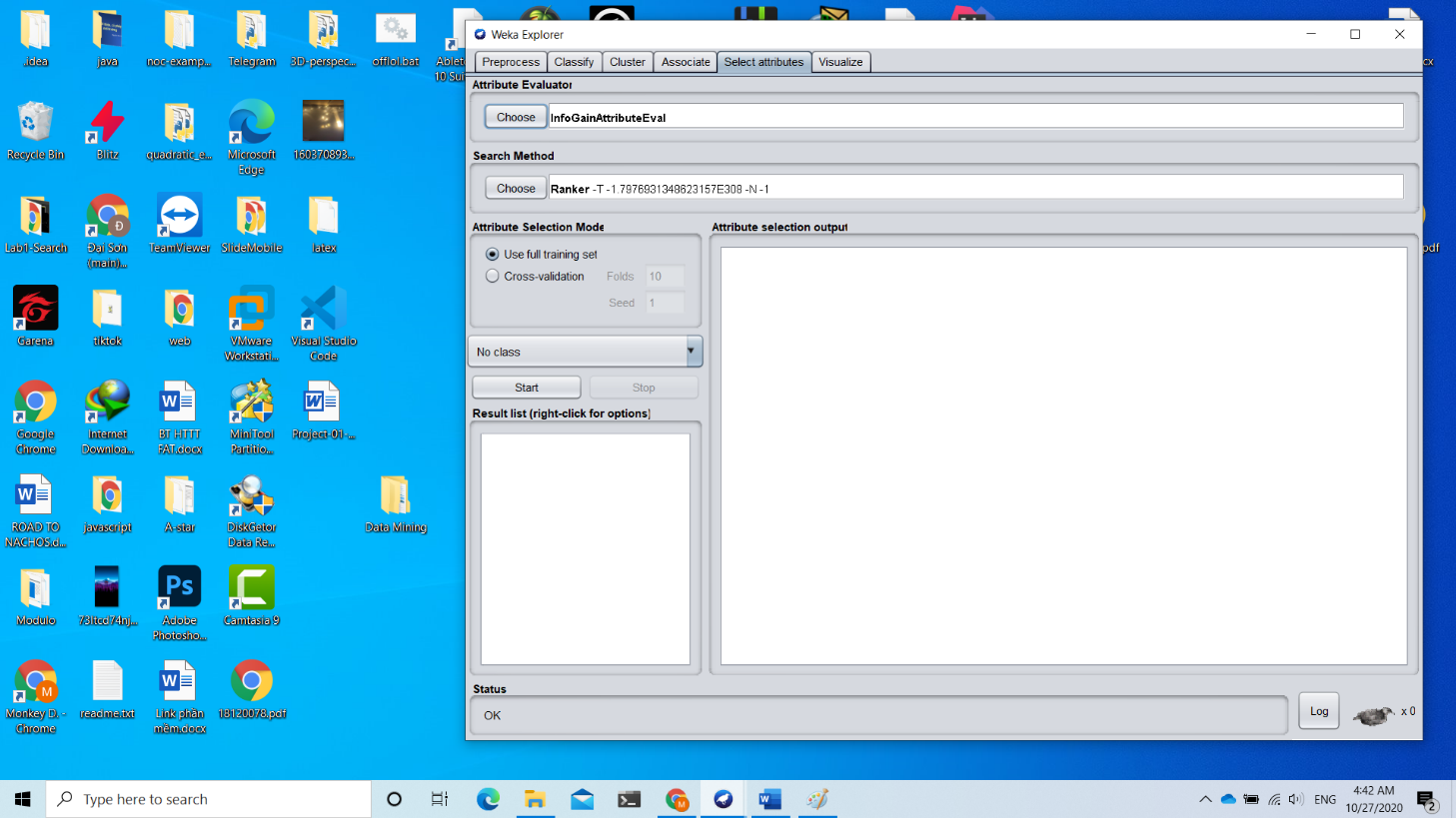
+ Negative Correlation(-1): 2 thuộc tính tỉ lệ nghịch với nhau

* + - *Information Gain Based Feature Selection (InfoGainAttributeEval):* Chọn các thuộc tính có độ đo Information Gain cao. Information Gain có giá trì từ 0 (không có thông tin) đến 1(chứa thông tin tối đa)
    - *Learner Based Feature Selection (WrapperSubsetEval):* Sử dụng 1 thuộc toán học mạnh để áp dụng trên 1 tập các thuộc tính. Đồ chính xác của giải thuật học trên tập thuộc tính nào được xấp xỉ nhờ cross-validation. Chọn tập các thuộc tính đưa ra kết quả học tốt nhất
    - *SymmetricalUncertAttributeEval:* Đánh giá giá trị của một tập hợp các thuộc tính bằng cách đo SU (Symmetrical Uncertainty) đối với thuộc tính phân lớp

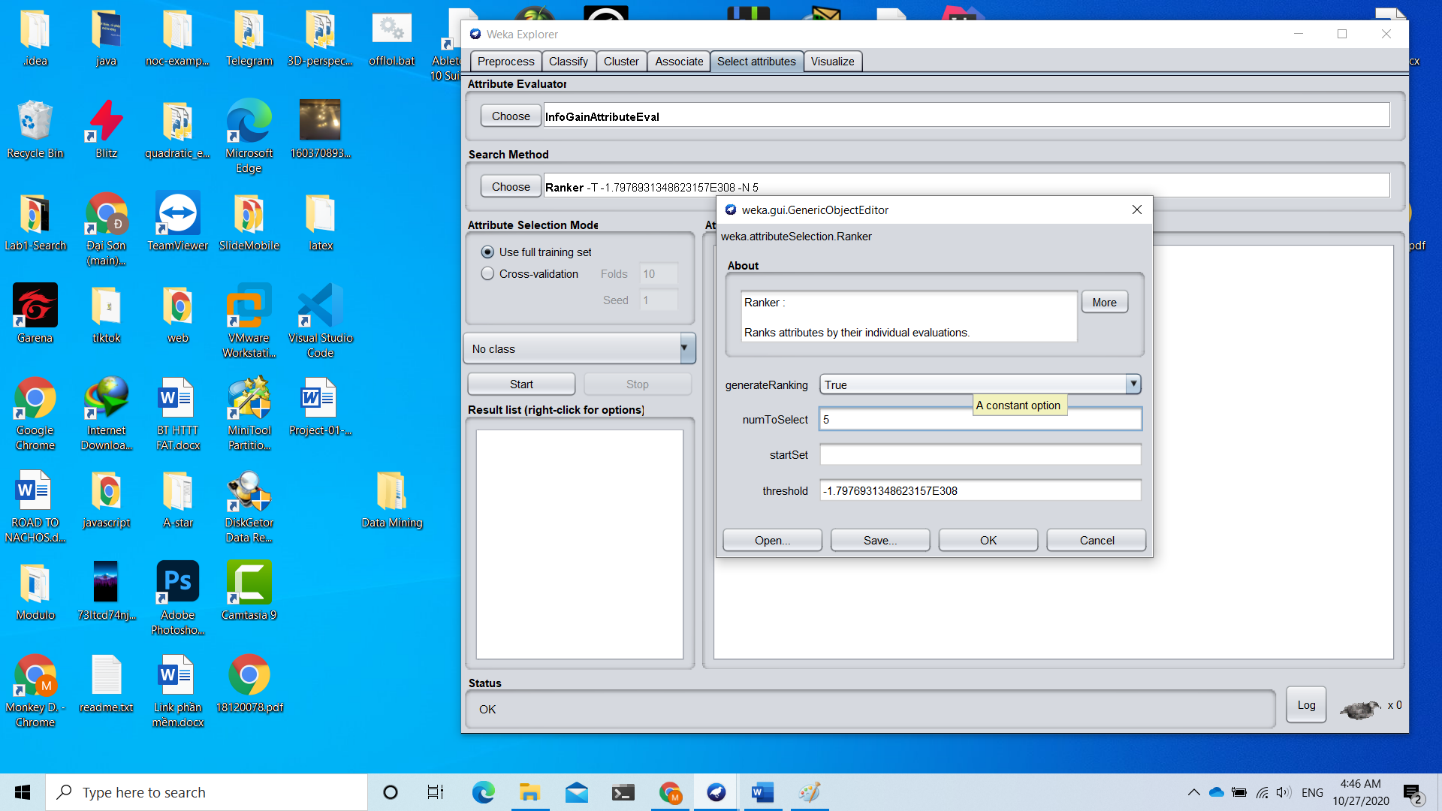
*+ Search Method:*

* + - *Best First:* Tìm kiếm không gian của các tập con thuộc tính bằng thuật toán Beam search. Việc đặt số lượng nút không cải tiến liên tiếp cho phép kiểm soát mức độ thực hiện quay lui. Best First có thể bắt đầu với tập thuộc tính trống và tìm kiếm hoặc bắt đầu với tập hợp đầy đủ các thuộc tính và tìm kiếm ngược lại, hoặc bắt đầu tại bất kỳ điểm nào và tìm kiếm theo cả hai hướng (bằng cách xem xét tất cả các lần thêm và xóa thuộc tính đơn lẻ có thể có tại một điểm nhất định ).
    - *GreedyStepwise*: Thực hiện tìm kiếm leo đồi tham lam theo 2 hướng tiến và lùi thông qua không gian của các tập con thuộc tính. Có thể bắt đầu với không có / tất cả các thuộc tính hoặc từ một điểm tùy ý trong không gian. Dừng khi việc thêm / xóa bất kỳ thuộc tính còn lại nào dẫn đến giảm đánh giá.
    - *Ranker:* xếp hạng các thuộc tính theo đánh giá cá nhân của nó.

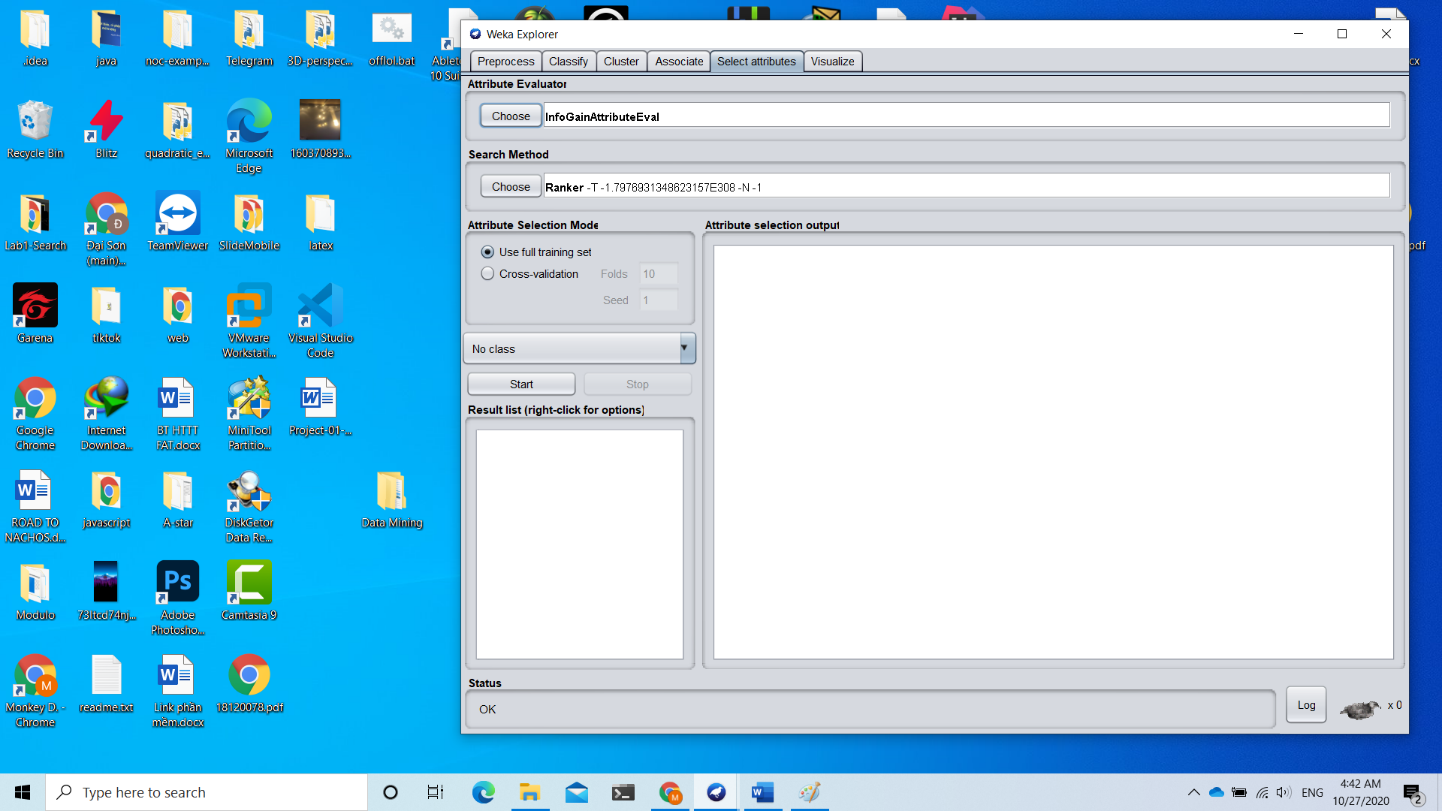
1. *Cần sử dụng bộ lọc nào để chọn ra 5 thuộc tính có tương quan cao nhất với thuộc tính lớp? Mô tả các bước làm, kèm theo hình chụp từng bước và kết quả cuối cùng.*
   * + Cần sử dụng phương pháp Ranker với 1 đồ độ đánh giá đơn thuộc tính (Ví dụ như Information Gain). Đến xếp hạng cho 5 thuộc tính có độ tương quan cao nhất với thuộc tính lớp.
     + Các bước thực hiện:
       - Chọn InfoGainAttributeEval tại mục Attribute Evaluator, chọn mục Ranker tại mục Search Method

**

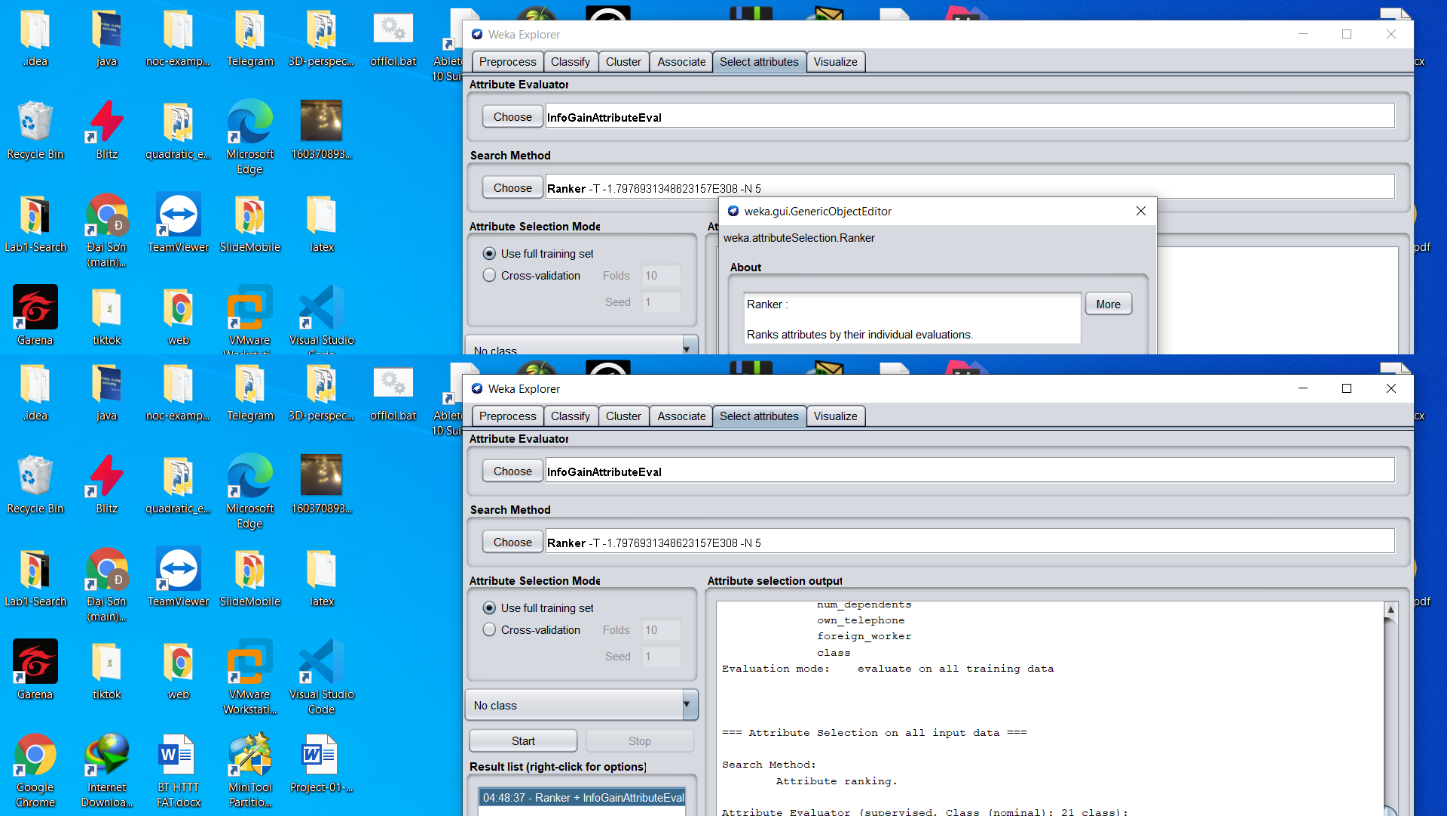
* + - * Nhấp chuột vào ô Search Method, đổi numToSelect thành 5 và nhấp OK

**

* + - * Tại mục Atribute Selection Mode chọn Use full training set và nhấp Start

**

* + - * Kết quả chọn lọc thuộc tính

**

### Cài đặt tiền xử lý dữ liệu:

* 1. **Liệt kê các cột bị thiếu dữ liệu**
  2. **Đếm số dòng bị thiếu dữ liệu**
  3. **Điền giá trị bị thiếu**
  4. **Xóa các dòng bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước**
  5. **Xóa các cột bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước**
  6. **Xóa các mẫu bị trùng lặp**
  7. **Chuẩn hóa một thuộc tính numeric bằng phương pháp min-mã và Z-score**
  8. **Tính giá trị biểu thức thuộc tính:**

## IV. Đánh giá đồ án

### 1. Mức độ hoàn thành của các thành viên

|  |  |  |
| --- | --- | --- |
| **MSSV** | **Mức độ hoàn thành công việc** | **Đóng góp** |
| 18120078 |  |  |
| 18120253 |  |  |

### 2. Mức độ hoàn thành đồ án:

## VII. Nguồn tham khảo