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:

### 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:**

### 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