**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**NHóm HTP**: Nguyễn Hoàng Hiệp – 15520214

lê hữu phúc – 15520639

nguyễn trường thịnh - 15520844



CHỦ ĐỀ: Dữ Liệu Điều Tra Dân Số

MÔN HỌC: Khai Thác Dữ Liệu

--------------------------------------------

Giảng viên hướng dẫn : - **TS.CAO THỊ NHẠN**

* **Th.s Vũ Minh sang**

# Lời cảm ơn

*Lời đầu tiên, nhóm em xin được gởi lời cảm ơn đặc biệt đến Cô -Ts Nguyễn Thị Kim Phụng và Thầy Th.s Vũ Minh Sang - Giảng viên phụ trách môn khai thác dữ liệu – trường đại học Công Nghệ Thông Tin – Đại Học Quốc Gia Tp Hồ Chí Minh .*

*Trong thời gian nhóm em làm đồ án , tụi em đã nhận được nhiều sự giúp đỡ từ cô và thầy. Cô/Thầy đã cung cấp đầy đủ kiến thức, chỉ bảo và đóng góp những ý kiến quý báu giúp tụi em có thể hoàn thành được đồ án môn học của mình một cách tốt nhất.*

*Xuất phát từ mục đích học tập, tìm hiểu xâu hơn các kiến thức về khai thác dự liệu, cũng như tìm hiểu kỹ về quy trình nghiệp vụ của lên ý tưởng, xây dựng một Model khai thác dữ liệu. Nhóm chúng em đã thực hiện đồ án “Khai thác dữ liệu từ tập dữ liều điều tra dân số năm 1994.”. Trong quá trình thực hiện đồ án, dữa trên kiến thức được Thầy cung cấp qua các buổi học lý thuyết cũng như thực hành trên lớp, kết hợp với việc tự tìm hiểu những công cụ và kiến thức mới, nhóm đã cố gắng thực hiện đồ án một cách tốt nhất .Tuy nhiên, đồ án còn chưa được hoàn hiện và có nhiều sai sót.*

*Nhóm rất mong nhận được sự góp ý từ cô nhằm rút ra những kinh nghiệm quý báu và hoàn thiện vốn kiến thức để nhóm có thể hoàn thành những đồ án, dự án khác trong tương lai .*

***Nhóm chúng em xin chân thành cảm ơn quý cô/thầy!***

***Lời Nhận Xét Của Giảng Viên***

**…………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………….………………………………………………………………………………..…………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………..…………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………..……………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………**

Mục lục

[*Lời cảm ơn* 1](#_Toc533498028)

[Mục lục 3](#_Toc533498029)

[Chương 1: Lý do chọn Dataset và giới thiệu tổng quan Dataset 5](#_Toc533498030)

[1.1 Giới thiệu tổng quan Dataset 6](#_Toc533498031)

[1.1.1 Nguồn dữ liệu sử dụng 6](#_Toc533498032)

[1.1.1.1 Hướng dẫn tải dataset thực hiện trong đồ án và các dataset khác của nhà cung cấp 6](#_Toc533498033)

[1.1.2 Mô tả chi tiết dữ liệu 6](#_Toc533498034)

[1.1.3 Mô tả mục đích bài toán 7](#_Toc533498035)

[1.1.4 Tiền xử lý dữ liệu 7](#_Toc533498036)

[1.1.4.1 Làm sạch dữ liệu 7](#_Toc533498037)

[1.1.4.2 Tích hợp dữ liệu 7](#_Toc533498038)

[1.1.4.3 Thu giảm, rút gọn dữ liệu 9](#_Toc533498039)

[1.1.4.4 Dữ liệu sau tiền sử lý dữ liệu 9](#_Toc533498040)

[1.1.4.5 Chuẩn bị dữ liệu để huấn luyện và kiểm thử 10](#_Toc533498041)

[1.1.4.5.1 Train set 10](#_Toc533498042)

[1.1.4.5.2 Test set 11](#_Toc533498043)

[1.1.5 Mô tả chi tiết các thuộc tính trong dataset 13](#_Toc533498044)

[1.1.6 Giới thiệu các công cụ được sử dụng trong đồ án 15](#_Toc533498045)

[1.1.6.1 Tổng quan Weka 3.8 15](#_Toc533498046)

[1.1.6.2 Giới thiệu Python3 16](#_Toc533498047)

[1.1.6.3 Giới thiệu IDE PyCharm 17](#_Toc533498048)

[Chương 2: Thuật toán khai thác dữ liệu sử dụng (OLAP) 18](#_Toc533498049)

[2.1 Sử dụng thuật toán phân lớp dữ liệu(cây quyết định - J48) 19](#_Toc533498050)

[2.1.1 Tổng quan về thuận toán phân lớp dựa trên cây quyết định 19](#_Toc533498051)

[2.1.2 Lý do chọn thuật toán 19](#_Toc533498052)

[2.1.3 Quá trình thực hiện 19](#_Toc533498053)

[2.1.3.1 Train - set 19](#_Toc533498054)

[2.1.3.1.1 Kết quả đạt được 21](#_Toc533498055)

[2.1.3.2 Test – set 21](#_Toc533498056)

[2.1.3.2.1 Kết quả đạt được 23](#_Toc533498057)

[2.1.4 So sánh đánh giá 24](#_Toc533498058)

[2.2 Sử dụng thuật toán NavieBayes(mô hình xác xuất) 24](#_Toc533498059)

[2.2.1 Lý do chọn thuật toán 24](#_Toc533498060)

[2.2.2 Train – set 24](#_Toc533498061)

[2.2.3 Test – set 25](#_Toc533498062)

[2.3 Sử dụng thuật toán K-means 26](#_Toc533498063)

[2.3.1 Lý do chọn thuật toán 26](#_Toc533498064)

[2.3.2 Tập train – set 27](#_Toc533498065)

[2.3.3 Tập Test – set 32](#_Toc533498066)

[2.3.4 Kết quả thu được 33](#_Toc533498067)

[Chương 3: Kết luận 35](#_Toc533498068)

[3.1 Kết quả đạt được 36](#_Toc533498069)

[3.2 Những hạn chế 36](#_Toc533498070)

[3.3 Hướng phát triển trong tương lai 36](#_Toc533498071)

[3.4 Bảng phân công nhiệm vụ trong nhóm 37](#_Toc533498072)

[3.5 Tài liệu tham khảo 38](#_Toc533498073)

1. Lý do chọn Dataset và giới thiệu tổng quan Dataset

***G****iới thiệu tổng quan về dataset, lý do hình thành dự án, nguồn dữ liệu thực hiện. Trình bày chi tiết các kỹ thuật xử lý tiền xử lý xử dụng và lý do xử dụng kỹ thuật đó.*

* 1. Giới thiệu tổng quan Dataset
     1. Nguồn dữ liệu sử dụng

Nguồn dữ liệu là tập hợp các bản ghi kết quả điều tra dân số năm 1994 do Barry Becker thực hiện từ cuộc tổng điều tra dân số.

Nguồn dữ liệu được thu thập từ: <https://archive.ics.uci.edu/ml/machine-learning-databases/adult/>

* + - 1. Hướng dẫn tải dataset thực hiện trong đồ án và các dataset khác của nhà cung cấp

Link xem mô tả chi tiết các thuộc tính: <https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names>.

Link tải dataset:

<https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data>

Link tải dataset 2:

<https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.test>

Bao gồm dữ liệu kết quả khảo sát của năm 1994 với định dạng.data .

* + 1. Mô tả chi tiết dữ liệu

Dữ liệu Adult là tập hợp các bảng ghi thực hiện bởi Barry Becker từ cuộc tổng điều tra dân số năm 1994.

* + 1. Mô tả mục đích bài toán

Từ dữ liệu, thông qua việc phân tích dựa trên các thuật toán khai thác dữ liệu. Dưa ra tập luật để dự đoán xem một người có đặt tính như thế nào thì có mức lương hơn 50.000$ một năm.

* + 1. Tiền xử lý dữ liệu

Kỹ thuật tiền sử xử lý được chọn

* + - 1. Làm sạch dữ liệu

Dữ liệu được trính xuất làm sạch để tiến hành khai thác dữ liệu.

Dữ liệu có những chổ trống, nguyên nhân do người thực hiện khảo sát không điền.

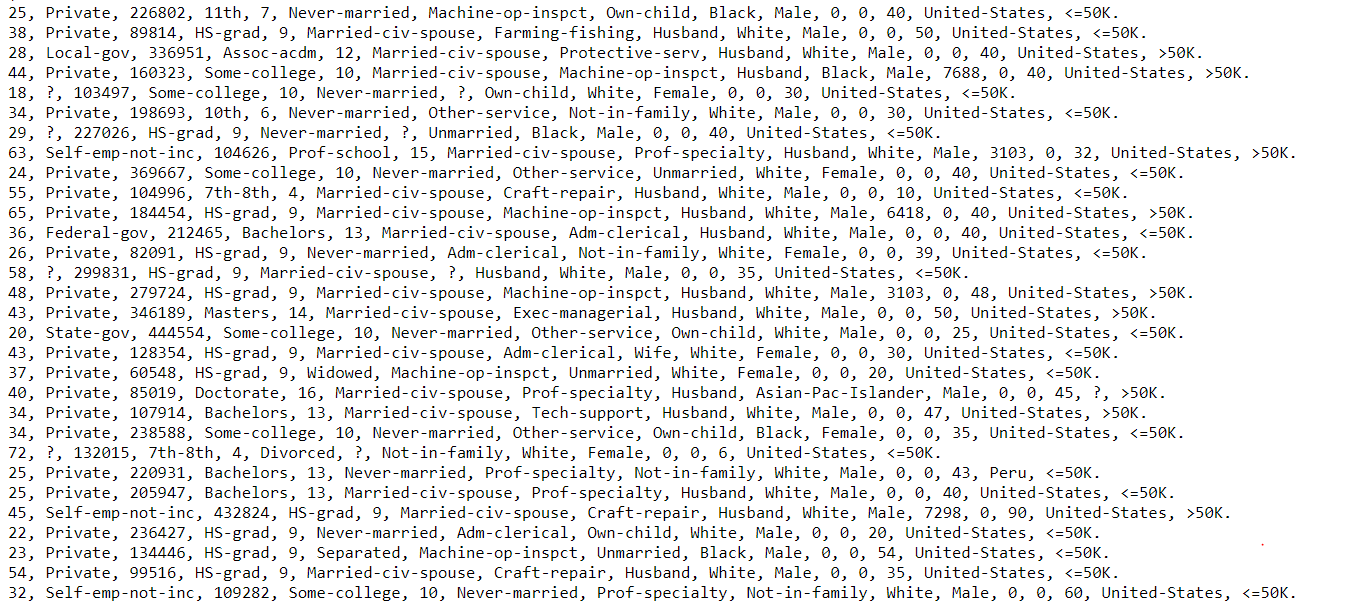
Giải pháp: Ngăn dự liệu bị thiếu, bắt buộc điền các thuộc tính cần thiết.

* + - 1. Tích hợp dữ liệu

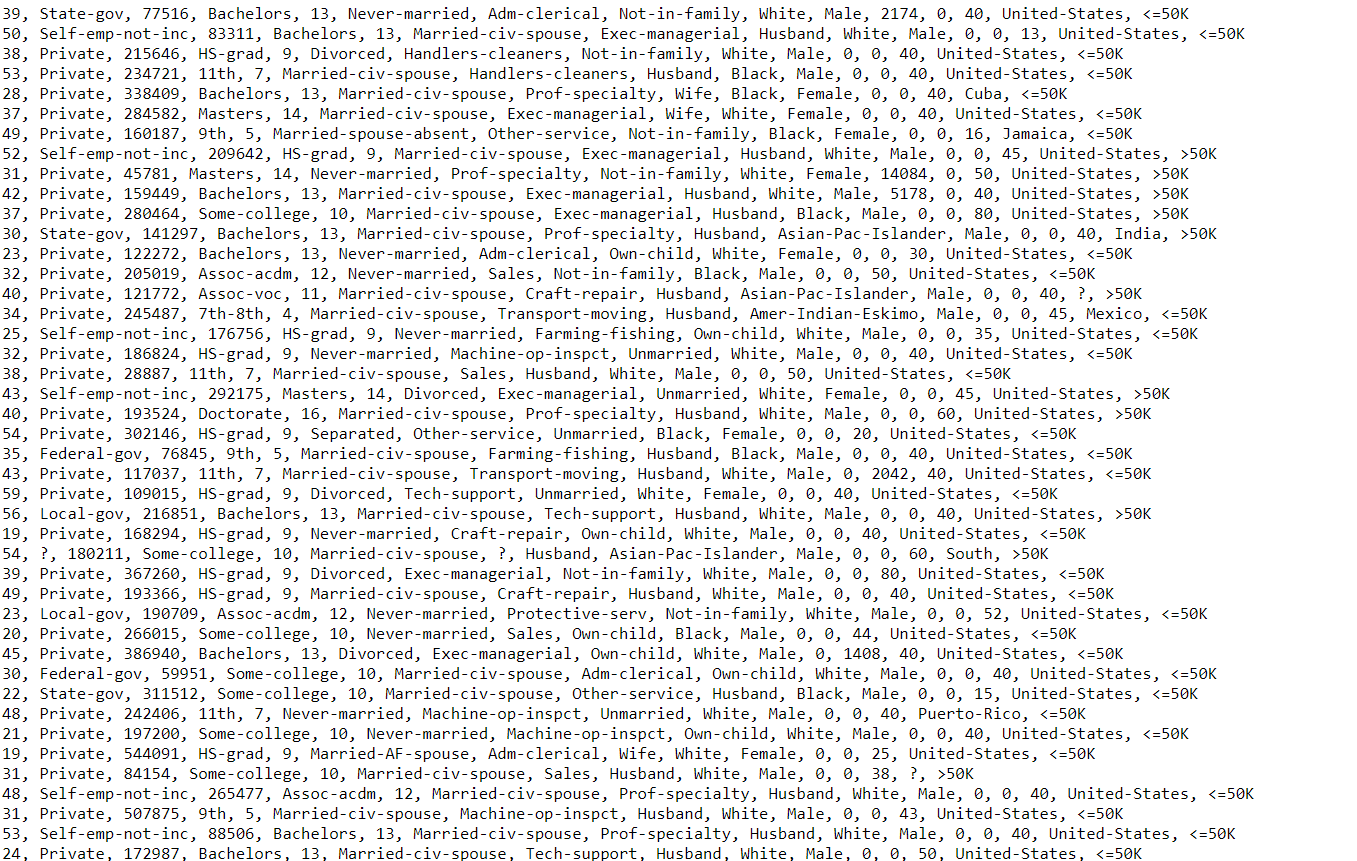
Dữ liệu tìm được từ 2 file riêng biệt nhưng cấu trúc tương đương giữa các nguồn

Nên ta tiến hành tích hợp 2 file dữ liệu vào một kho duy nhất để hiệu quả khai thác dữ liệu cao hơn(tăng độ chính xác,..).

Dữ liệu mẫu từ nguồn thứ nhất:



Dữ liệu mẫu từ nguồn thứ 2:

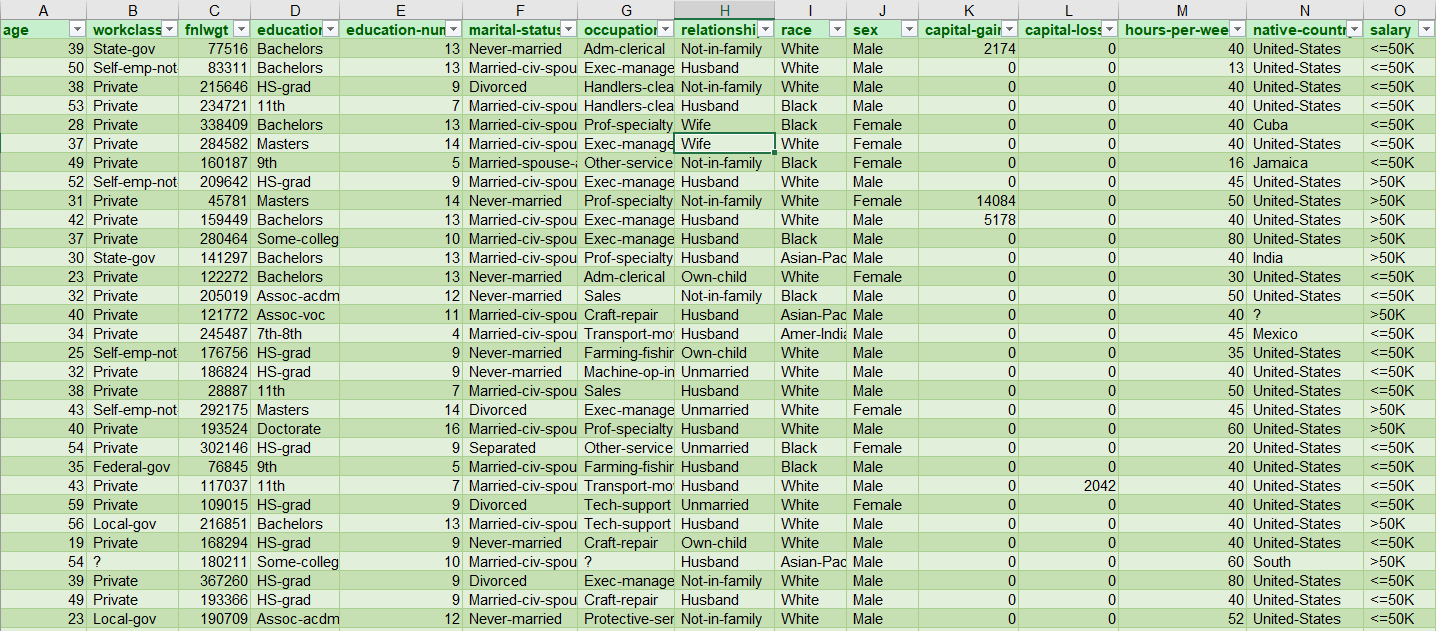


* + - 1. Thu giảm, rút gọn dữ liệu

Tập hợp các bản ghi sạch được chiết xuất dữ liệu dựa trên những điều kiện sau:

* + Tuổi phải lớn hơn 16 và nhỏ hơn 90
  + Trọng lượng cuối cùng(final weight) [19214:1184622]
  + AGI lớn hơn 100
  + Số giờ làm một tuần [0;99].
    - 1. Dữ liệu sau tiền sử lý dữ liệu

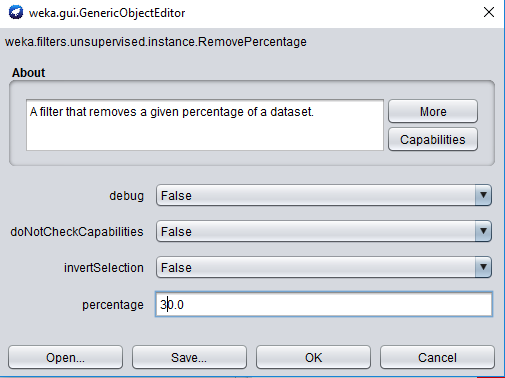
Dữ liệu sau khi trích xuất là còn 48842 người lớn và 15 thuộc tính:



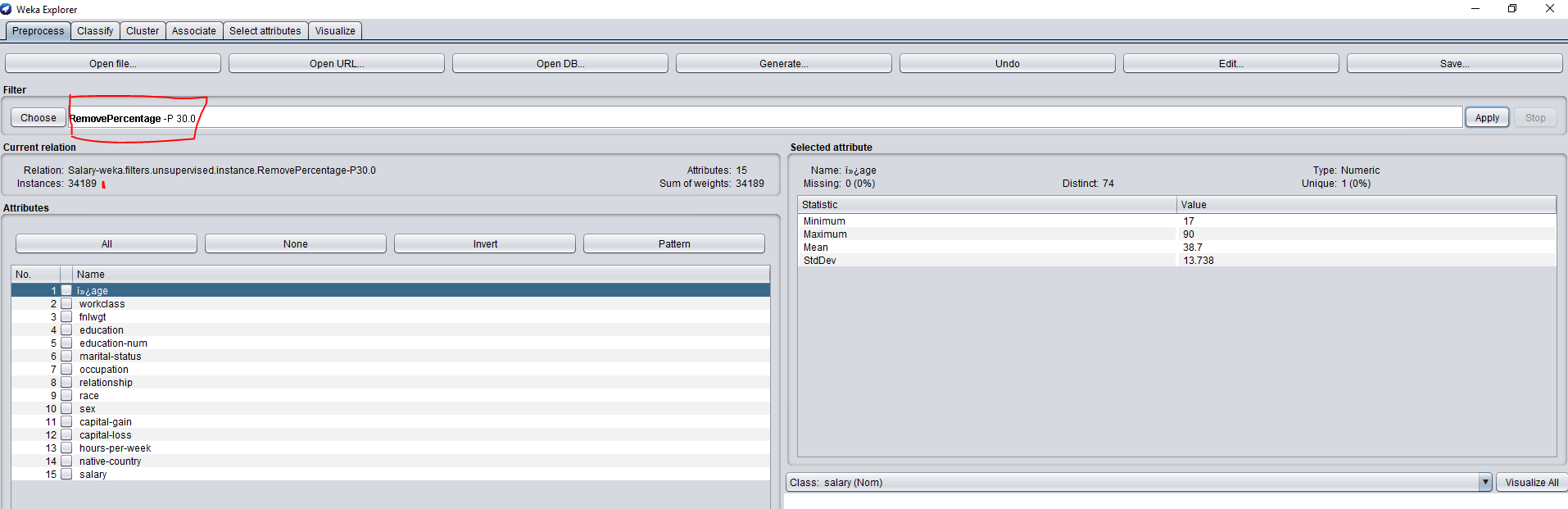
* + - 1. Chuẩn bị dữ liệu để huấn luyện và kiểm thử

Sử dụng RemovePercentage để chia tập huấn huyện và tập kiểm thử. Chia theo tỷ lệ 70:30.

* + - * 1. Train set

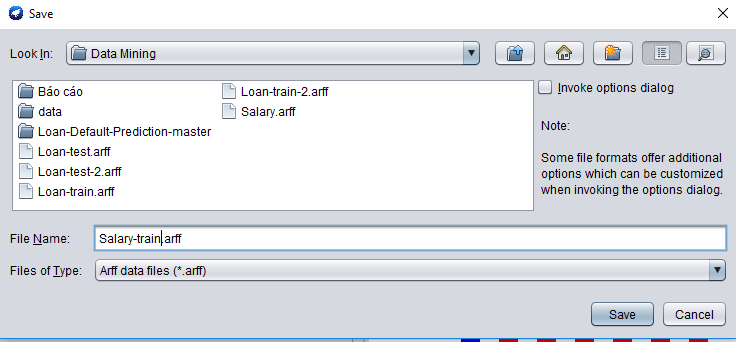


Nhấn OK -> nhấp Apply: thu được kết quả như sau:



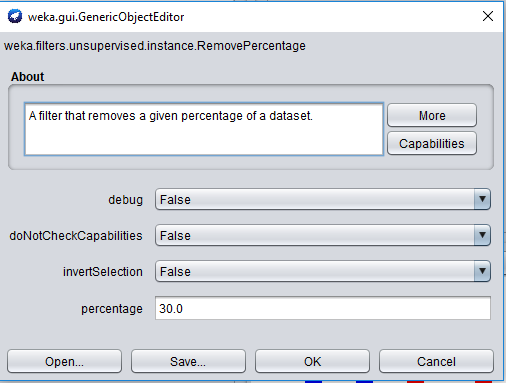
Tập dữ liệu huấn luyện có 34189 dòng.

Lưu lại tập huấn luyện:



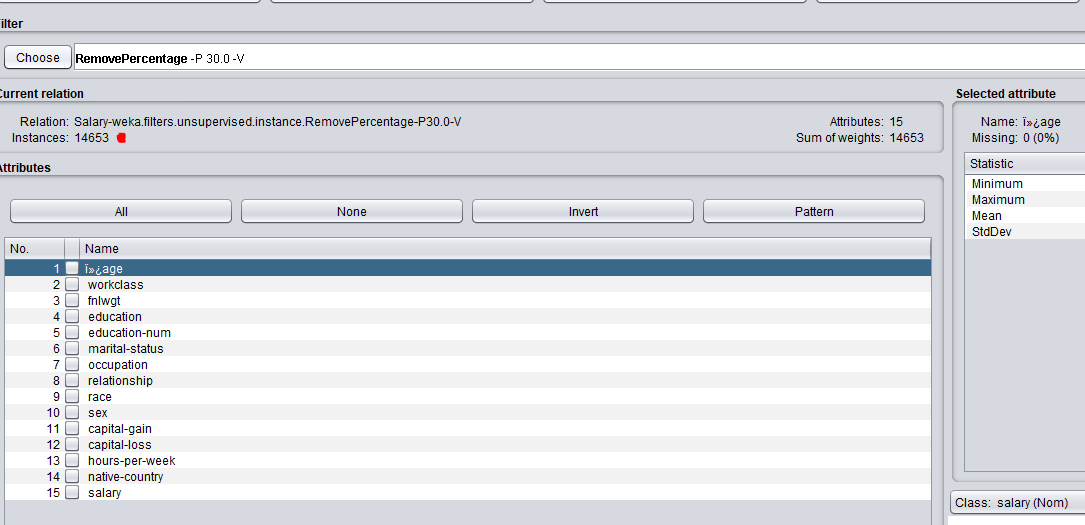
* + - * 1. Test set

Thực hiện tượng tự nhưng đổi invertSelection thành TRUE:

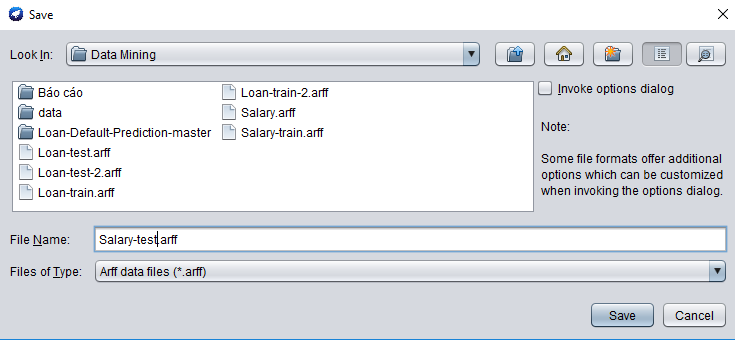


Nhấn OK -> Nhấn Apply: kết quả thu được

Tập kiểm thử có 14653 dòng:



Lưu lại tập kiểm thử.



**Như vậy ta đã có 2 tập dữ liệu để trainning là test như mong muốn:**

* + 1. Mô tả chi tiết các thuộc tính trong dataset

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Ý nghĩa** | **Giá trị** |
|  | Age | Tuổi | [17; 90] |
|  | workclass |  | { Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.  } |
|  | fnlwgt | Trọng lượng cuối cùng(final weight) | [19214; 1184622] |
|  | Education | Học vấn | { Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool  } |
|  | education-num |  |  |
|  | marital-status | Tình trạng hôn nhân | { Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.  } |
|  | occupation | Ngành nghề | { Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces  } |
|  | relationship | Mối quan hệ thân nhân | { Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.  } |
|  | race | Chủng tộc | { White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.  } |
|  | Sex | Giới tính | {Male; Female} |
|  | capital-gain | Vốn tăng | [0; 9999] |
|  | capital-loss | Vốn giảm | [0;4365] |
|  | hours-per-week | Số giờ làm 1 tuần | [0; 99] |
|  | native-country | Quê quán | { United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands  } |
|  | Salary | Lương/1năm (thuộc tính phân lớp) | {<=50K; >50K} |

* + 1. Giới thiệu các công cụ được sử dụng trong đồ án
* Phần mềm Weka 3.8
* Ngôn ngữ lập trình: Python3
* IDE sử dụng: PyCharm
  + - 1. Tổng quan Weka 3.8

Weka là một phần mềm khai thác dữ liệu. Thuộc dự án đại học Waikato – New Zealand.

Là phần mềm mã nguồn mỡ, xây dựng bằng java, theo kiến trúc hướng đối tượng.

Theo KDNuggets(2005): Weka là sản phẩm khai thác dữ liệu được sử dụng nhiều nhất và hiệu quả nhất trong năm 2005.

Những tính năng vượt trội của Weka:

* + Hỗ trợ nhiều thuận toán máy học và khai thác dữ liệu
  + Là mã nguồn mở
  + Độc lập với môi trường(do sử dụng JVM)
  + Dễ sử dụng, kiến trúc dạng thư viện dễ dàng cho việc xây dựng các ứng dụng thực nghiệm

Kiến trúc thư viện Weka:

* + Hơn 600 lớp
  + Tổ chức thành 10 gói Package, mỗi gói thực hiện một nhiệm vụ trong quá trình khai thác dữ liệu
  + Có thể sử dụng trực tiếp trên phân mềm hay dùng các lớp thư viện phát triển ứng dụng.

Trang chủ: <https://www.cs.waikato.ac.nz/ml/weka/>

Link download: <https://www.cs.waikato.ac.nz/ml/weka/downloading.html>

* + - 1. Giới thiệu Python3

Python là một ngôn ngữ lập trình bậc cao cho các mục đích lập trình đa năng, do Guido van Rossum tạo ra và lần đầu ra mắt vào năm 1991. Python được thiết kế với ưu điểm mạnh là dễ đọc, dễ học và dễ nhớ. Python là ngôn ngữ có hình thức rất sáng sủa, cấu trúc rõ ràng, thuận tiện cho người mới học lập trình. Cấu trúc của Python còn cho phép người sử dụng viết mã lệnh với số lần gõ phím tối thiểu. Vào tháng 7 năm 2018, Van Rossum đã từ chức Leader trong cộng đồng ngôn ngữ Python sau 30 năm lãnh đạo.

Python hoàn toàn tạo kiểu động và dùng cơ chế cấp phát bộ nhớ tự động; do vậy nó tương tự như Perl, Ruby, Scheme, Smalltalk, và Tcl. Python được phát triển trong một dự án mã mở, do tổ chức phi lợi nhuận Python Software Foundation quản lý.

Ban đầu, Python được phát triển để chạy trên nền Unix. Nhưng rồi theo thời gian, nó đã "bành trướng" sang mọi hệ điều hành từ MS-DOS đến Mac OS, OS/2, Windows, Linux và các hệ điều hành khác thuộc họ Unix. Mặc dù sự phát triển của Python có sự đóng góp của rất nhiều cá nhân, nhưng Guido van Rossum hiện nay vẫn là tác giả chủ yếu của Python. Ông giữ vai trò chủ chốt trong việc quyết định hướng phát triển của Python.

Sau đây là các đặc điểm của **Python**:

* Ngữ pháp đơn giản, dễ đọc.
* Vừa hướng thủ tục (procedural-oriented), vừa hướng đối tượng (object-oriented)
* Hỗ trợ module và hỗ trợ gói (package)
* Xử lý lỗi bằng ngoại lệ (Exception)
* Kiểu dữ liệu động ở mức cao.
* Có các bộ thư viện chuẩn và các module ngoài, đáp ứng tất cả các nhu cầu lập trình.
* Có thể nhúng vào ứng dụng như một giao tiếp kịch bản (scripting interface).

Python 3, còn gọi là Python 3000 hoặc Py3K: Dòng 3.x sẽ không hoàn toàn tương thích với dòng 2.x, tuy vậy có công cụ hỗ trợ chuyển đổi từ các phiên bản 2.x sang 3.x. Nguyên tắc chủ đạo để phát triển Python 3.x là "bỏ cách làm việc cũ nhằm hạn chế trùng lặp về mặt chức năng của Python". Trong PEP (Python Enhancement Proposal) có mô tả chi tiết các thay đổi trong Python. Các đặc điểm mới của Python 3.0 sẽ được trình bày phần cuối bài này.

* + - 1. Giới thiệu IDE PyCharm

Professional là một trong những môi trường phát triển Python tốt nhất giúp tạo ra các ứng dụng Python, Django, và phát triển Web, một Python IDE thông minh trong việc phát hiện lỗi và sửa lỗi mã, chỉnh sửa, xử lý sự cố và nhiều hơn nữa.. Bộ biên tập mã thông minh của PyCharm cung cấp hỗ trợ cho Python, JavaScript, CoffeeScript, TypeScript, CSS, các ngôn ngữ mẫu phổ biến và hơn thế nữa.

* Tính năng, đặc điểm :
* Trình biên tập nâng cao Python
* Sử dụng màu sắc của các mã
* Phân tích chất lượng mã viết
* Gỡ lỗi Nhanh
* Khả năng chỉnh sửa và chỉnh sửa mã Django, JavaScript và HTML
* Hỗ trợ Google App Engine cho Python và nhiều khuôn khổ khác như Pyridam, CYTHON, WEB2PY và …
* Tích hợp với Hệ thống Kiểm soát Phiên bản (VCS)
* Khả năng tùy biến và cá nhân hóa giao diện người dùng

1. Thuật toán khai thác dữ liệu sử dụng

***C****hương này sẽ đi vào giới thiệu thuật toán khai thác dữ liệu được chọn, giải thích lý do chọn thuật toán và phù hợp với dữ liệu như thế nào.*

* 1. Sử dụng thuật toán phân lớp dữ liệu(cây quyết định - J48)
     1. Tổng quan về thuận toán phân lớp dựa trên cây quyết định

Cây quyết định là một cấu trúc phân cấp của các nút và các nhánh.

Có 3 loại nút trên cây:

* Nút gốc
* Nút nội bộ: Nút mang tên thuộc tính CSDL
* Nút lá: mang tên lớp

Nhánh mang giá trị thuộc tính

* + 1. Lý do chọn thuật toán

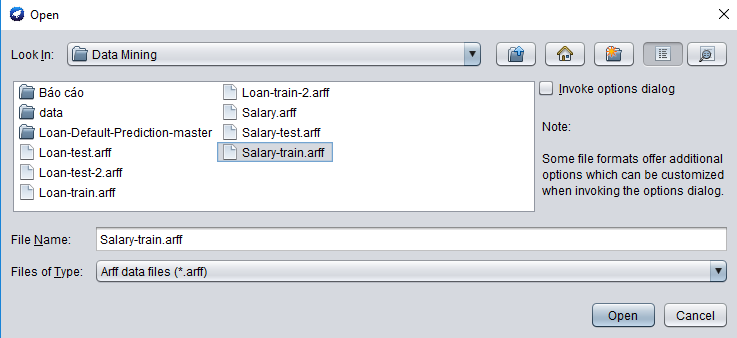
Lý do nhóm e bắt đầu với phân lớp dữ liệu dựa trên cây quyết định vì:

* Nó dễ dang để xây dựng
* Phân lớp mẫu mới nhanh
* Dễ dàng diễn giải cho các cây kích thước nhỏ
* Độ chính xác là khá tốt so với các thuật toán phân lớp khác trên tập dữ liệu
  + 1. Quá trình thực hiện

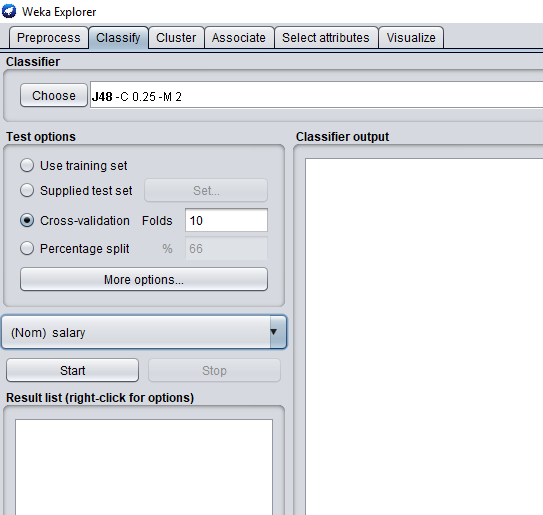
Sử dụng dữ liệu train – set và test – set lần lượt và thiết lập 10-fold cross validation với thuật toán J48

* + - 1. Train - set

**Mở dữ liệu train – set:**

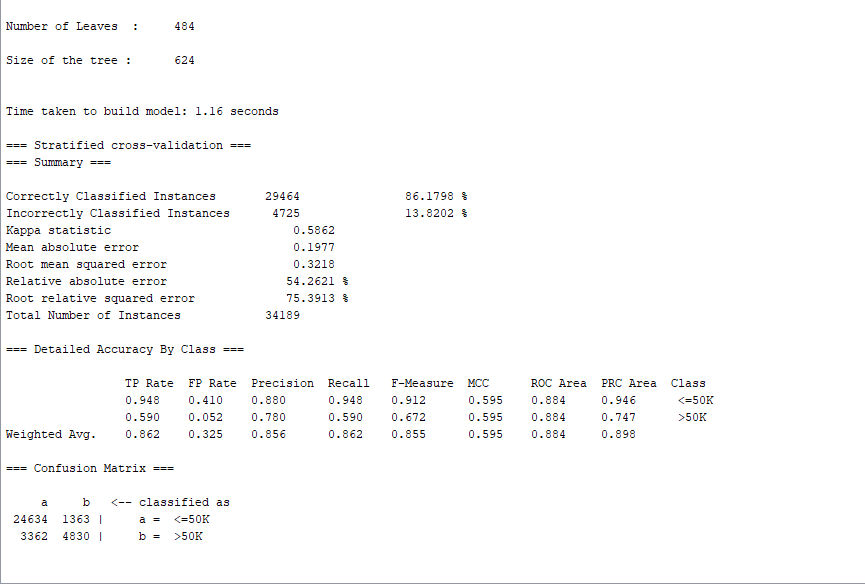


**Chuyển qua tag Classify:** Chọn thuật tán J48, chọn Test option là Cross-validation Folds =10 -> chọn thuộc tính phân lớp là salary:



Nhấn Start để tiền hành.

* + - * 1. Kết quả đạt được

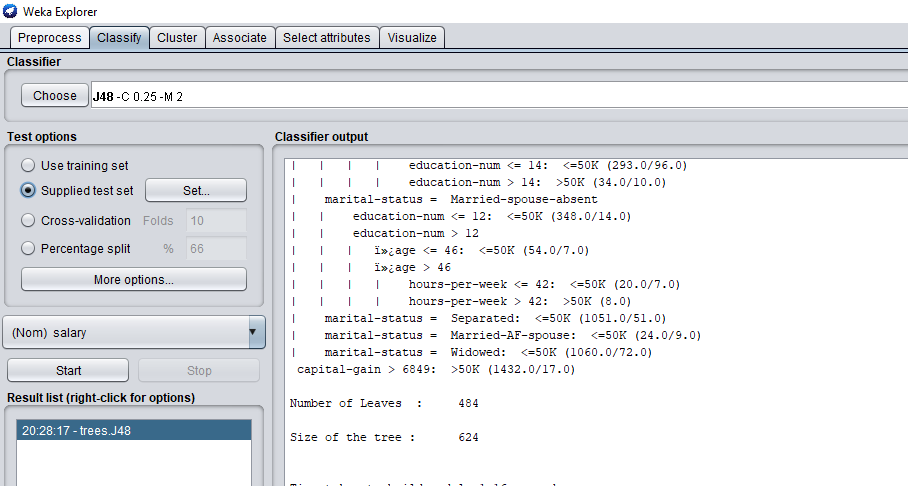


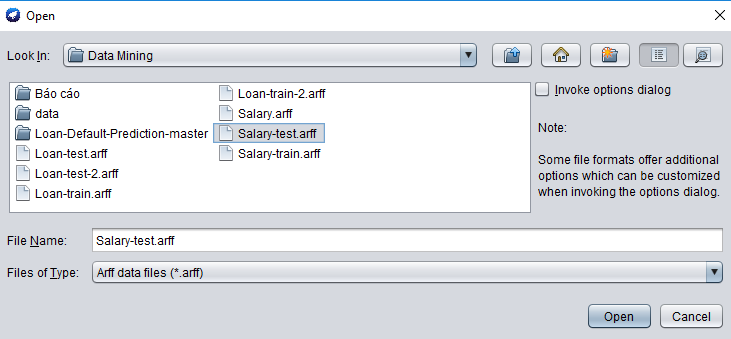
Mô hình cây quyết định này đạt độ chính xác là 86% và không chính xác là 14%.

* + - 1. Test – set

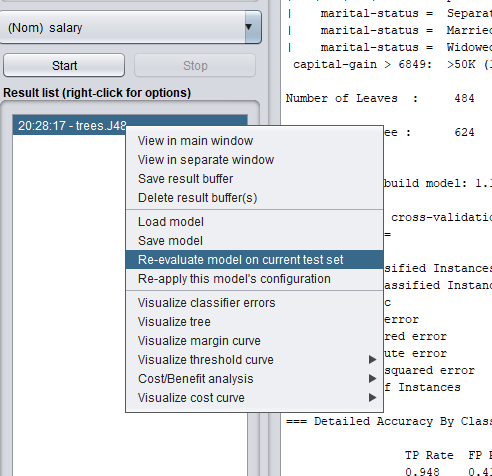
Thực hiện model trên với tập kiểm thử:

Chọn **Supplied test set** -> chọn **Set** -> chọn file **test - set**

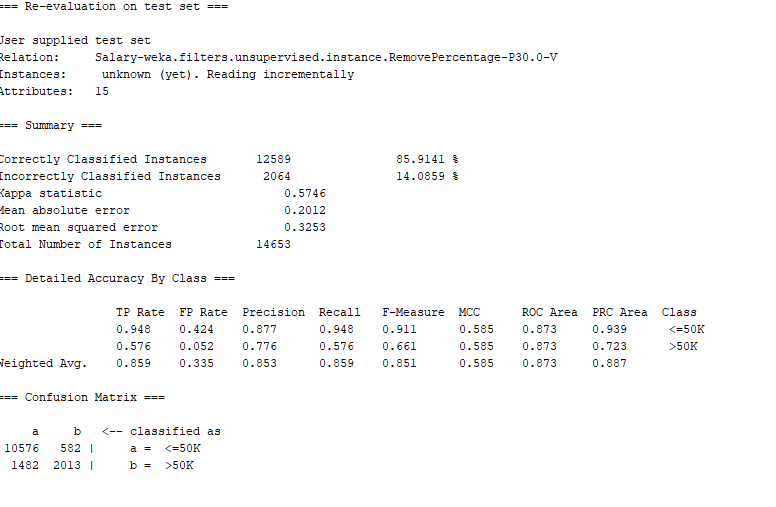




Chuột phải mode đã làm với tập train -> chọn Re-veluate model on curent test set



* + - * 1. Kết quả đạt được



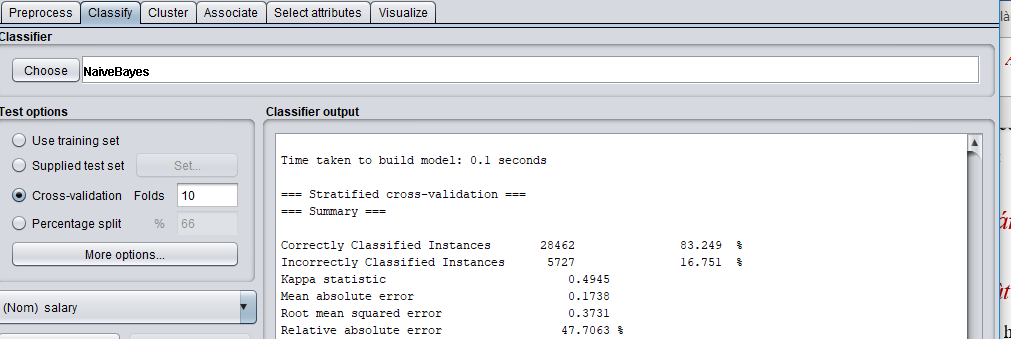
Nhận xét: Mô hình cây quyết định này đạt độ chính xác là 86% và không chính xác là 14% với tập test -set.

* + 1. So sánh đánh giá

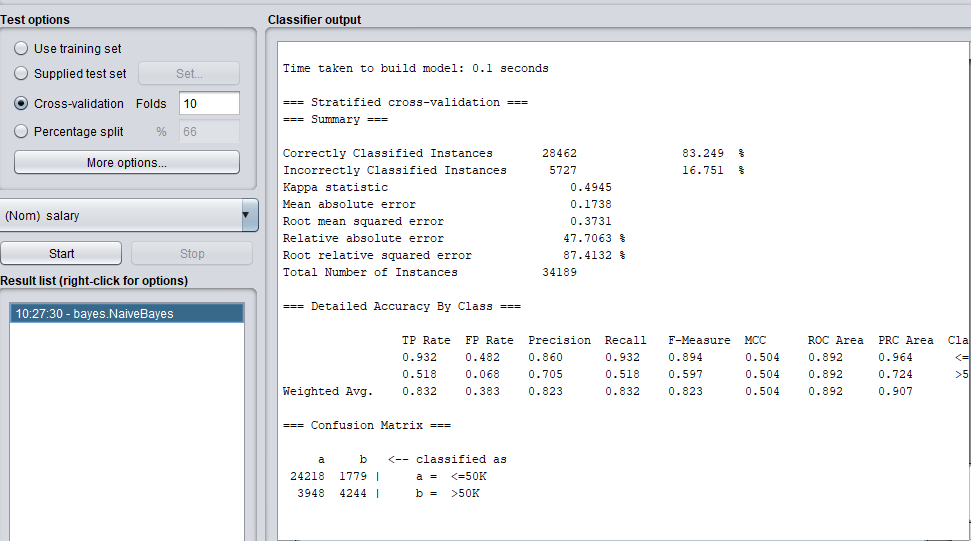
Như vậy ta nhận được kết quả từ 2 tập (train và test) là gần như nhau và đều đạt độ chính xác cao (86%).

* 1. Sử dụng thuật toán NavieBayes(mô hình xác xuất)
     1. Lý do chọn thuật toán
* Dễ dàng cài đặt và thực hiện
* Thời gian thi hành tương tự như cây quyết định
* Đạt kết quả tốt trong phần lớn các trường hợp
  + 1. Train – set

Thực hiện tương tự như với J48 nhưng chọn thuật toán NavieBayes



Kết quả thu được:



Đọc kết quả:

Correctly Classified Instances: số mẫu đã phân lớp đúng là 28462 tương đương ~83%.

Incorrectly Classified Instances: Số mẫu phân lớp sai là 5727 tương đương ~17%

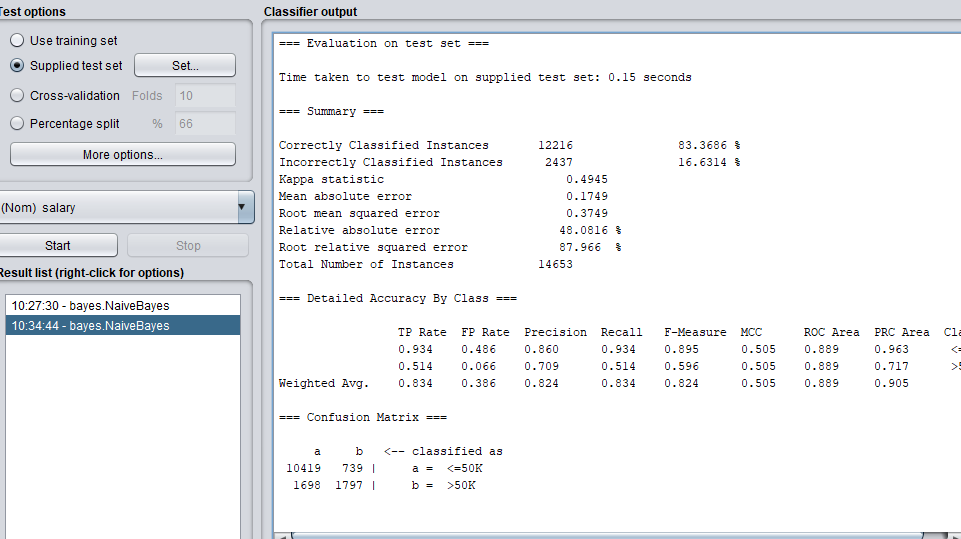
Detailed Accuracy By Class: cho ta biết độ chính xác của từng phân lớp.

Confusion Matrix: có 1779 thuộc lớp >50K nhưng lại bị gán vào <=50K

Và có 3948 mẫu thuộc lớp <=50K mà lại bị gán vào mẫu >50K

* + 1. Test – set

Thực hiện tương tự, thu được kết quả:



Đọc kết quả:

Correctly Classified Instances: số mẫu đã phân lớp đúng là 12216 tương đương ~83%.

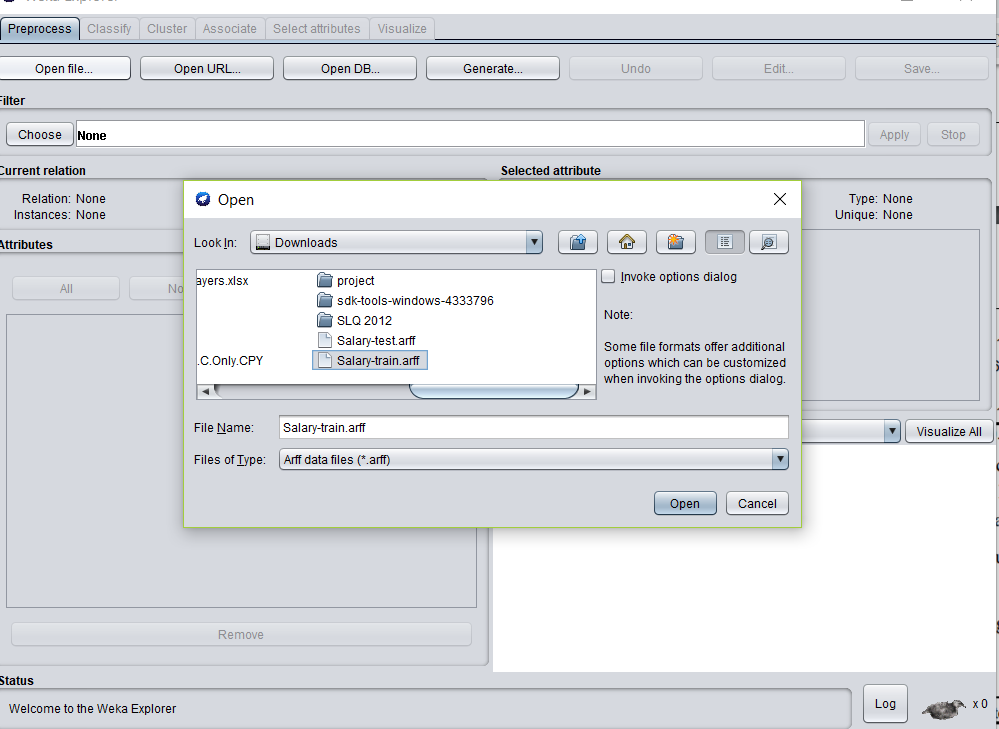
Incorrectly Classified Instances: Số mẫu phân lớp sai là 2437 tương đương ~17%

Detailed Accuracy By Class: cho ta biết độ chính xác của từng phân lớp.

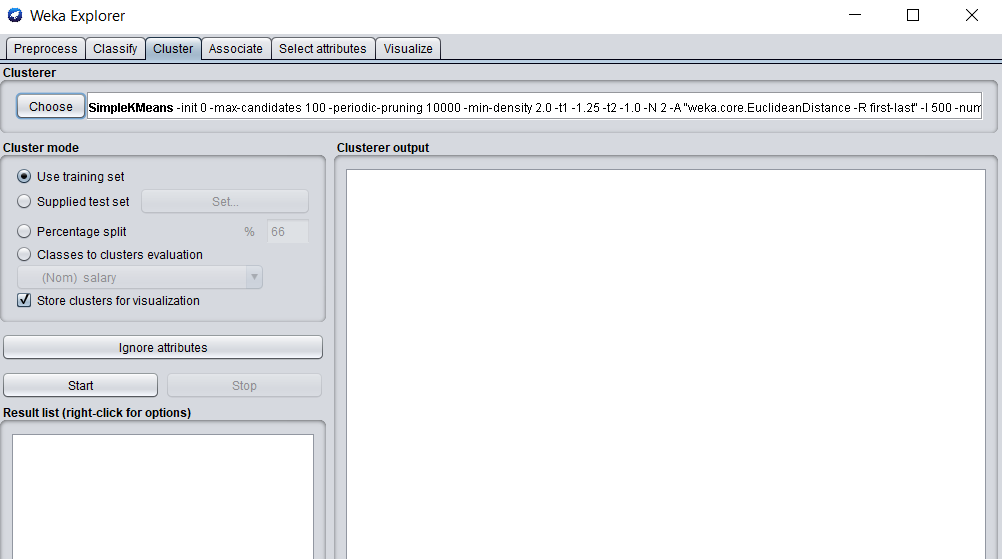
Confusion Matrix: có 739 thuộc lớp >50K nhưng lại bị gán vào <=50K

Và có 1698 mẫu thuộc lớp <=50K mà lại bị gán vào mẫu >50K

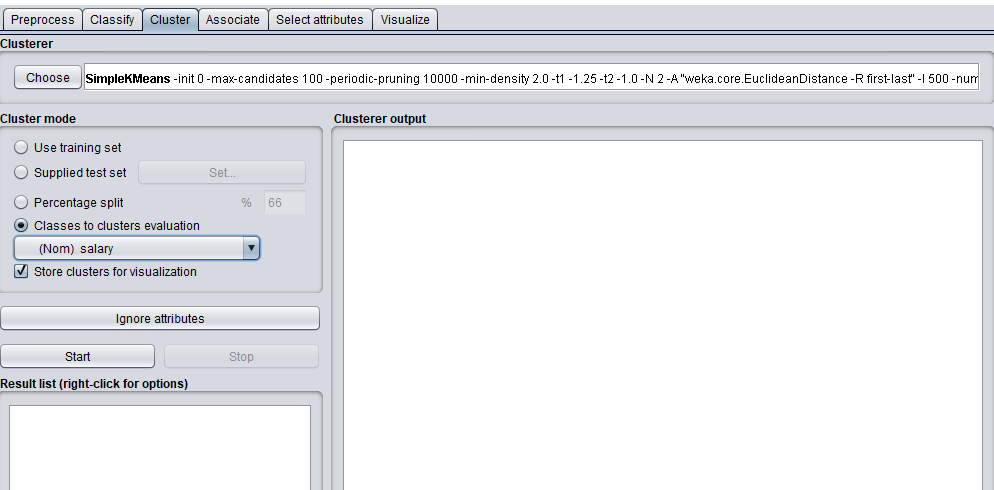
* 1. Sử dụng thuật toán K-means
     1. Lý do chọn thuật toán
* Đơn giản dễ hiểu, tương đối hiệu quả
* Các đối tượng tự động gán vào các nhóm
* Thường đạt tối ưu cục bộ
  + 1. Tập train – set



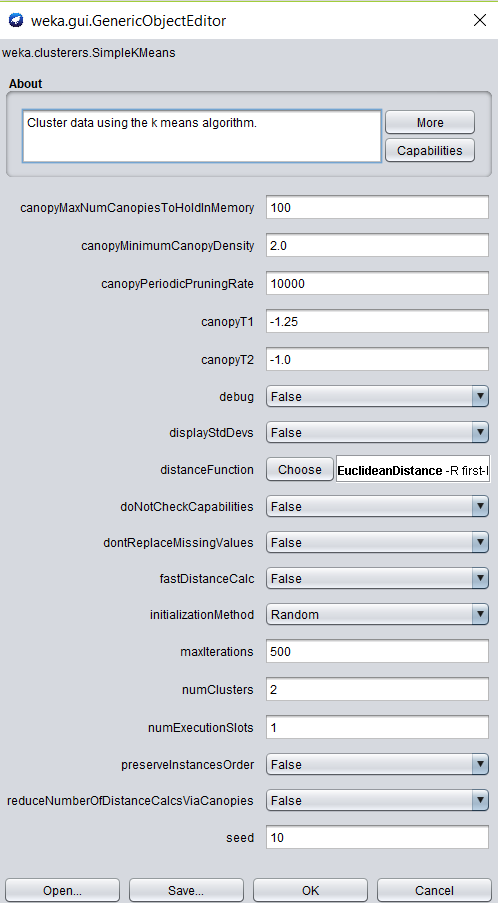
Vào Cluster 🡪 Choose 🡪 Simple K-Means



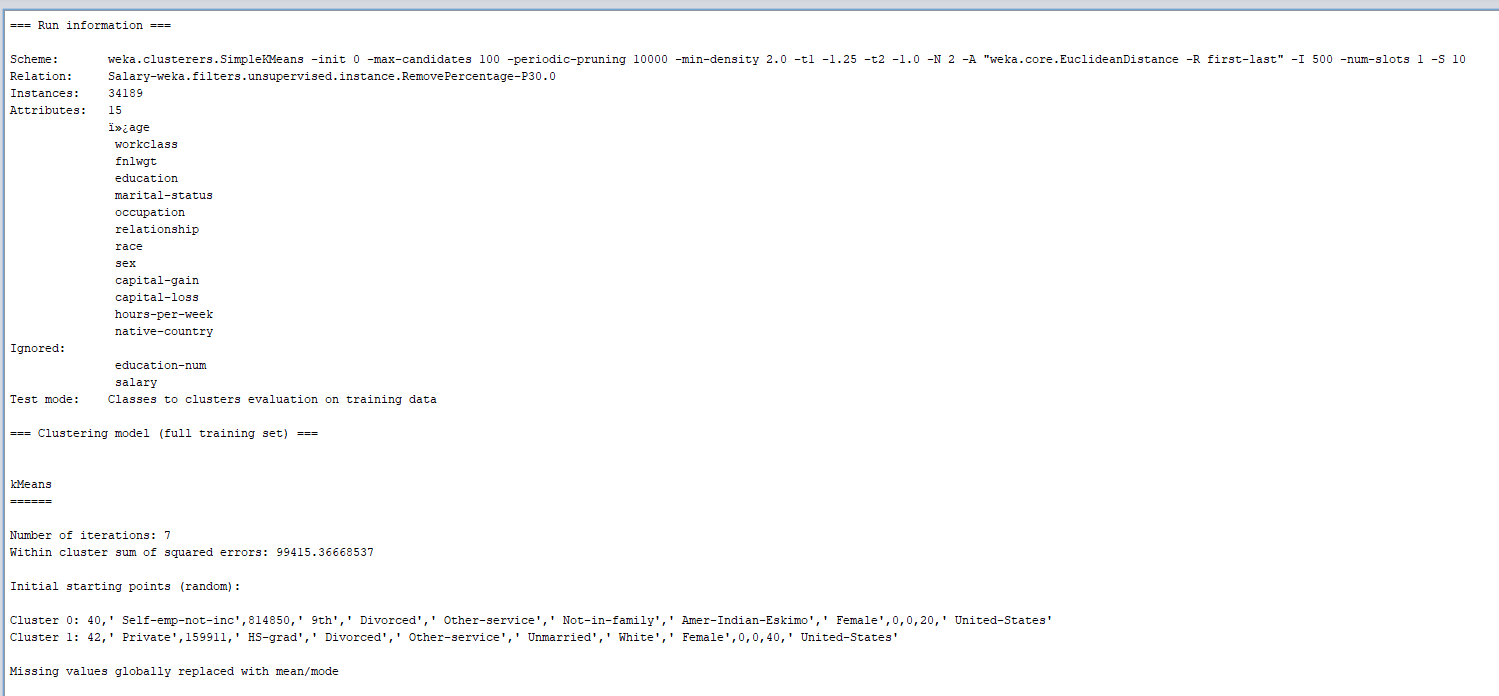
* Chọn Classes to clusters evaluatiton 🡪 (Nom) Salary

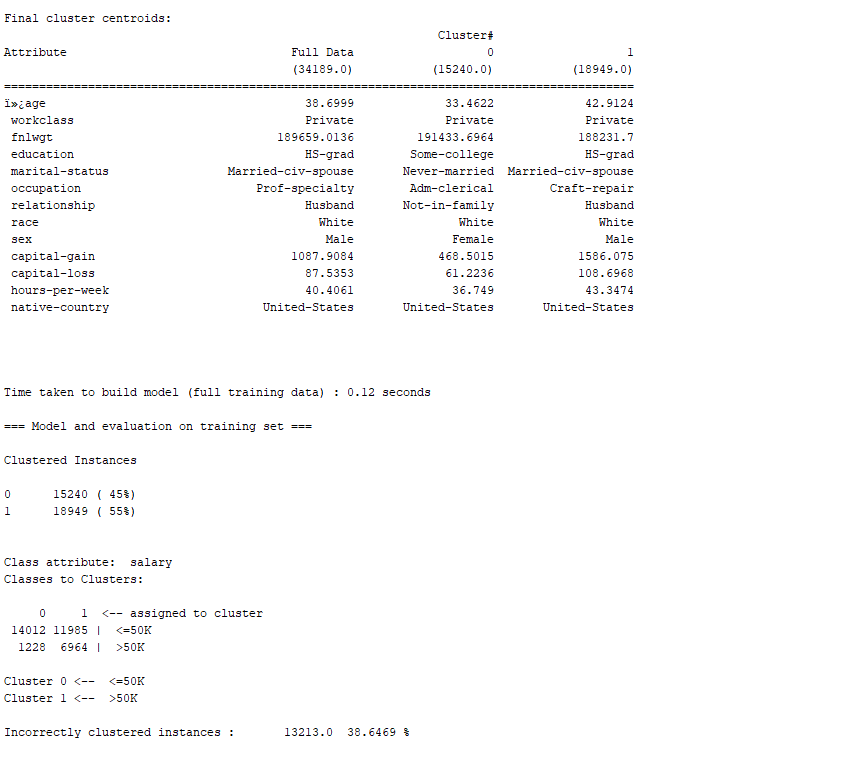


* Nhấn chuột trái vào SimpleKMeans trên thanh Choose 🡪 xuất hiện bảng weka.gui.GenericObjectEditor 🡪 dòng numCluster: 2



* Chọn Start 🡪 Kết quả phân tích





Phân tích kết quả:

* Mô hình gom nhóm được sử dụng: **SimpleKMeans**
* Kiểu kiểm tra: Thực hiện kiểm chứng trên tập huấn luyện **salarytrain.arff**
* Tổng số mẫu phân tích: **34189**
* Mối quan hệ: **Tiền lương (Salary).** Bao gồm các yếu tố: Age, Workclass, fnlwgt, Education, Marital – status, Occupation, Relationship, Sex, Capital – gain, Capital – loss, Hour – per – week, navite – country. Sẽ quyết định số tiền lương.
* Cluster chia thành 2 nhóm

**+** Nhóm 0 (Salary <= 50K): **15240** chiếm **45%**

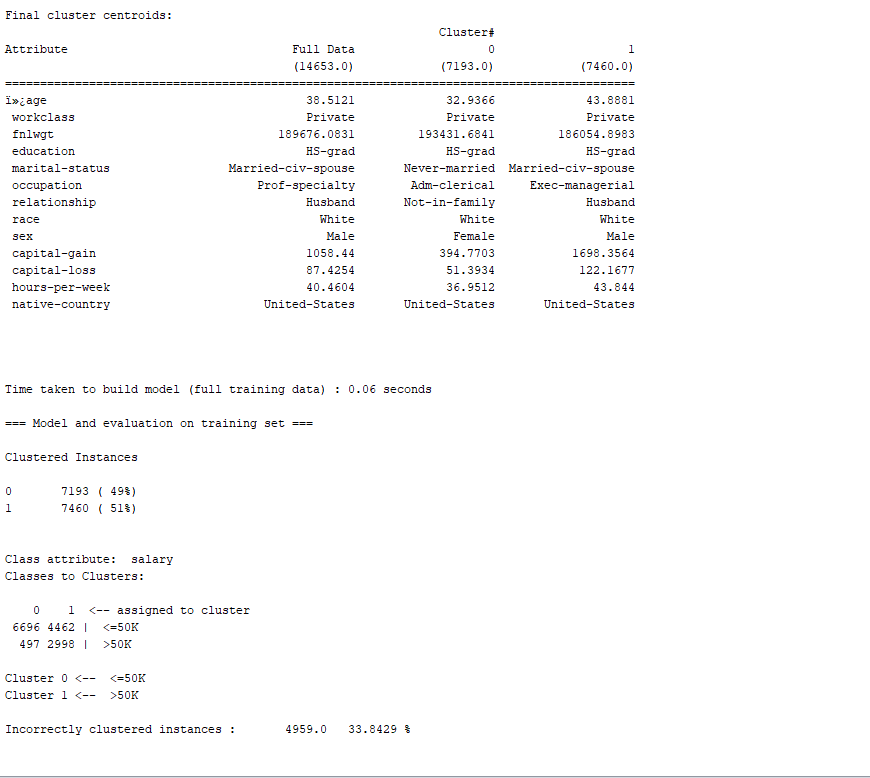
**+** Nhóm 1 (Salary > 50K): **18949** chiếm **55%**

Độ sai sót của phép phân tích này là **38.6469%**

* + 1. Tập Test – set

Thực hiện các bước như đã làm trên tập dữ liệu salarytrain.arff

**Kết quả phân tích trên weka:**



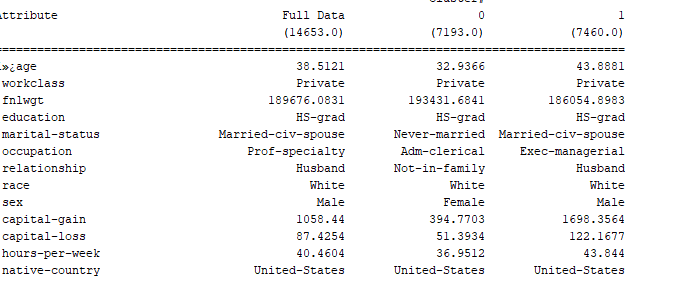
**Phân tích kết quả:**

* Mô hình gom nhóm được sử dụng: **SimpleKMeans**
* Kiểu kiểm tra: Thực hiện kiểm chứng trên tập huấn luyện **salarytrain.arff**
* Tổng số mẫu phân tích: **14653**
* Mối quan hệ: Tiền lương (Salary). Bao gồm các yếu tố: Age, Workclass, fnlwgt, Education, Marital – status, Occupation, Relationship, Sex, Capital – gain, Capital – loss, Hour – per – week, navite – country. Sẽ quyết định số tiền lương.
* Cluster chia thành 2 nhóm

**+** Nhóm 0 (Salary <= 50K): **7193** chiếm **49%**

**+** Nhóm 1 (Salary > 50K): **7460** chiếm **51%**

* Độ sai sót của phép phân tích này là **33.8429%**
  + 1. Kết quả thu được



* Bảng trên là giá trị trung bình của các thuộc tính thuộc 2 nhóm cluster. Cụ thể :

**+ Cluster 0 (Salary <= 50K)**:

. age: 32.9366

. workclass: Private

. fnlwgt: 193431.6841

. education: HS – grad

. marital – status: Never – married

. occupation: Adm – Clerical

. relationship: Not – in – family

. race: White

. sex: Female

. Capital – gain: 394.7703

. Capital – loss: 51.3934

. hours – per – week: 36.9512

. native – country: US

**+ Cluster 1 (Salary > 50K)**:

. age: 43.8881

. workclass: Private

. fnlwgt: 186054.8983

. education: HS – grad

. marital – status: Married – civ - spouse

. occupation: Exec - managerial

. relationship: Husband

. race: White

. sex: Male

. Capital – gain: 1698.3564

. Capital – loss: 122.1677

. hours – per – week: 43.844

. native – country: US

1. Kết luận

***Ở*** *chương này, nêu ra các kết quả đạt được, những hạn chế còn tồn tại sau khi thực hiện đồ án. Nêu ra hướng phát triển trong tương lai. Bảng phân công công việc và tài liệu tham khảo khi thực hiện đồ án.*

* 1. Kết quả đạt được

Trong kỳ học vừa qua , nhóm đã tìm hiểu và vận dụng kiến thức về khai thác dữ liệu và đạt được các kết quả như sau :

* Nắm rõ các khái niệm cơ bản về khai thác dữ liệu
* Năm rõ các thuật toán và kỹ thuật tiền xử ký dữ liệu
* Năm vững kiến thức và có thể vận dụng, xử dụng các thuận toán đã học thực hiện khai thác dữ liệu “Điều tra dân số”
* Trang bị kiến thức về tập phổ biến và luật kết hợp, khai thác chuỗi tuần tự, phân lớp dữ liệu, gom nhóm dữ liệu
* Biết được ưu nhược điểm của từng thuật toán
  1. Những hạn chế

Do thời gian hạn ngắn cộng với khối lượng công việc nhiều nên trong quá trình thực hiện đồ án nhóm còn gặp phải một số vấn đề :

* Áp dụng tiền sử lý dữ liệu chưa được tối ưu.
* Do dữ liệu có nhiều thuộc tính liên tục dẫn đến việc sử dụng thuật toán chưa được tối ưu.
  1. Hướng phát triển trong tương lai
* Áp dụng các thuật toán khác có thể cho ra kết quả tốt hơn
* Áp dụng Deep learning để khám phá để dự đoán ngững người có lương hơn 50.000$ một năm được chính xác hơn
* Áp dụng nhiều thuật toán khác nhau để tao ra nhiều tính năng
* Áp dụng vào thực tiễn, tạo ra thị trường ngách cho những sản phẩm cao cấp.
  1. Bảng phân công nhiệm vụ trong nhóm

|  |  |  |  |
| --- | --- | --- | --- |
| **Công việc** | **Nguyễn Hoàng Hiệp** | **Nguyễn trường thịnh** | **Lê Hữu Phúc** |
| Chọn dataset |  |  |  |
| Nghiên cứu tính cần thiết để khai thác dữ liệu |  |  |  |
| Mô tả bài toán |  |  |  |
| Tiền xủ lý dữ liệu |  |  |  |
| Chọn thuật toán |  |  |  |
| Thực hiện dữ liệu với thuật toán đã chọn |  |  |  |
| Giải thích kết quả thu được |  |  |  |
| So sánh, đánh giá các kết quả |  |  |  |
| Viết báo cáo |  |  |  |

* 1. Tài liệu tham khảo
* Slide bài giảng của cô Ts Nguyễn Thị Kim Phụng , thầy Th.s Vũ Minh Sang
* Sách:
  + Jiawei Han, Micheline Kamber, and Jian Pei, Data Mining Concepts and Techniques, 3 edition, Morgan Kaufmann Publishers, 2011.
* Video:
  + <https://www.youtube.com/watch?v=IJJN-Li0aIY&list=PL67CJL04EcjN4hIkgZT3dgyxETTXNOiZb&index=74>
  + <https://www.youtube.com/watch?v=UVzdfJugZiI>