

---

# KHAI PHÁ DỮ LIỆU: PHÂN TÍCH XẾP LOẠI TỐT NGHIỆP VÀ CƠ HỘI VIỆC LÀM CỦA SINH VIÊN SỬ DỤNG KỸ THUẬT PHÂN LỚP DATA MINING: ANALYSIS OF FINAL GPA AND EMPLOYMENT OPPORTUNITIES OF GRADUATES USING CLASSIFICATION TECHNIQUES

LÊ QUỐC TIẾN<sup>1\*</sup>, ĐẶNG HOÀNG ANH<sup>2</sup>

<sup>1</sup>Trường Đại học Hàng hải Việt Nam

<sup>2</sup>Phòng Tổ chức - Hành chính, Trường Đại học Hàng hải Việt Nam

\*Email liên hệ: letien@vimaru.edu.vn

## Tóm tắt

Khai phá dữ liệu ngày càng được áp dụng rộng rãi trong nhiều lĩnh vực, bao gồm cả giáo dục. Các công cụ khai phá dữ liệu được sử dụng để phân tích nguồn dữ liệu khổng lồ, nhằm thu được các thông tin, tri thức có giá trị. Phân lớp, một kỹ thuật rất quan trọng trong khai phá dữ liệu, giúp dự đoán các xu hướng từ dữ liệu có sẵn. Với các trường đại học, việc phân tích và dự đoán kết quả học tập cũng như cơ hội việc làm của sinh viên sau khi ra trường luôn được đặt lên hàng đầu. Bài viết này trình bày về việc xây dựng mô hình phân lớp với thuật toán C4.5 trên dữ liệu của sinh viên đã tốt nghiệp khoa Công nghệ thông tin nhằm phân tích, đánh giá mối quan hệ giữa kết quả học tập và cơ hội việc làm, từ đó hỗ trợ cho việc định hướng học tập và nghề nghiệp cho sinh viên trong tương lai.

**Từ khóa:** Khai phá dữ liệu, phân lớp, sinh viên tốt nghiệp, xếp loại tốt nghiệp, việc làm.

## Abstract

Data mining has become popular in many fields, including education. Data mining tools are used to extract valuable information and knowledge from big data sources. Classification which is an important technique in data mining makes the movement forecastable from available data. For universities, the priority is always placed on the analysis and forecast of students' study results and employment opportunities after graduation. This article discusses about building classification model with C4.5 algorithm, utilizing the data source from the Faculty of Information Technology graduates. Its objectives are to analyse and assess the relationship between study results and employment opportunities, then facilitate the study and career orientation for students in the future.

**Keywords:** Data mining, classification, graduate students, final GPA, careers.

## 1. Đặt vấn đề

Hiện nay, bên cạnh vấn đề tuyển sinh đầu vào, số lượng - chất lượng đầu ra cùng cơ hội việc làm và lựa chọn ngành nghề của các sinh viên sau khi tốt nghiệp luôn là mối quan tâm hàng đầu của các trường đại học nói riêng và toàn xã hội nói chung. Phân tích các dữ liệu về kết quả toàn khóa học của sinh viên, đưa ra những dự đoán về cơ hội việc làm để từ đó có những điều chỉnh, định hướng kịp thời trong quá trình đào tạo là mục tiêu quan trọng của mỗi trường đại học. Điều này hoàn toàn khả thi, nếu các trường có thể tận dụng được nguồn dữ liệu lớn của sinh viên, và áp dụng các kỹ thuật khai phá dữ liệu một cách phù hợp.

**Khai phá dữ liệu** (Data Mining) là quá trình sắp xếp, tính toán một số lượng lớn các tập dữ liệu để xác định các mẫu và tạo lập các mối quan hệ dữ liệu, nhằm giải quyết những bài toán thông qua phân tích dữ liệu. Nói cách khác, khai phá dữ liệu giúp trích xuất các ra các thông tin, tri thức có ích từ những nguồn dữ liệu khổng lồ. Các công cụ khai phá dữ liệu cho phép các đơn vị, tổ chức dự đoán những xu hướng phát triển trong tương lai. Khai phá dữ liệu là giai đoạn phân tích trong quá trình khám phá dữ liệu tri thức (Knowledge Discovery in Database - KDD). Quá trình này bao gồm các bước: Lựa chọn dữ liệu (Selection), tiền xử lý (Pre-processing), khai phá dữ liệu, biến đổi dữ liệu (Transformation) và đánh giá dữ liệu (Interpretation/evaluation). Một số các kỹ thuật và thuật toán được sử dụng trong quá trình khai phá dữ liệu như: phân lớp, phân cụm, luật kết hợp, cây quyết định,...

## 2. Các nghiên cứu liên quan

Trong những năm gần đây, khai phá dữ liệu cùng với các kỹ thuật như phân lớp, luật kết hợp ngày càng được sử dụng rộng rãi trong nhiều lĩnh vực, bao gồm cả giáo dục. Rất nhiều những công trình nghiên cứu hay các bài báo khoa học ứng dụng khai phá dữ liệu đối với dữ liệu sinh viên nhằm phân tích, dự đoán khả năng hay kết quả học tập của các sinh viên đại học. Fadl Elsid và Eltahir [1] đã áp dụng kỹ thuật phân lớp với thuật toán C4.5 (J48) trên cơ sở dữ liệu sinh viên nhằm đánh giá hiệu quả của thuật toán, đồng thời dự đoán kết quả học tập của sinh viên. Trong khi đó, K. Sumathi và nhóm nghiên cứu [2] cũng áp dụng phương pháp khai phá dữ liệu phân lớp với cây quyết định để phân tích vị trí việc làm của sinh viên dựa trên khả năng học tập.

### 3. Khai phá dữ liệu

Quá trình khai phá dữ liệu bao gồm các nhóm công việc sau: Phát hiện bất thường (Anomaly detection) - phát hiện các bản ghi dữ liệu bất thường, ngoại lệ hoặc sai lệch; khai phá luật kết hợp (Association rule) - tìm ra các mối quan hệ dữ liệu giữa các biến; phân cụm (Clustering) - phát hiện các nhóm và cấu trúc tương đồng trong dữ liệu, mà không sử dụng các cấu trúc dữ liệu đã biết; phân lớp (Classification) - tổng quát hóa cấu trúc đã biết để áp dụng cho dữ liệu mới; hồi quy (Regression) - cố gắng tìm ra một hàm nhằm mô hình hóa dữ liệu với ít lỗi nhất, để ước tính mối quan hệ giữa dữ liệu hay tập dữ liệu; tóm tắt hóa (Summarization) - cung cấp một biểu diễn nhỏ của tập dữ liệu, gồm tạo báo cáo và trực quan hóa.

Ba loại mô hình trong khai phá dữ liệu bao gồm: mô hình mô tả (Descriptive modeling), mô hình dự đoán (Predictive modeling) và mô hình đề xuất (Prescriptive modeling). Tương ứng với mỗi loại mô hình, có những phương pháp và kỹ thuật khai phá dữ liệu khác nhau, tùy thuộc vào yêu cầu của từng bài toán cụ thể, mục đích xây dựng mô hình, cũng như đặc tính của dữ liệu. Mô hình mô tả cung cấp các thông tin trong quá khứ có giá trị, tiết lộ những điểm tương đồng hoặc nhóm chung trong dữ liệu. Những kỹ thuật điển hình ở mô hình này: phân cụm - nhóm các dữ liệu có tính tương tự với nhau, sử dụng các thuật toán như K-means; luật kết hợp - phát hiện các mối quan hệ của dữ liệu, sử dụng các thuật toán như Apriori. Mô hình dự đoán đi sâu phân tích các dữ kiện trong tương lai, hoặc ước tính, dự báo các xu hướng sẽ xảy ra. Các kỹ thuật quan trọng ở mô hình này: phân lớp - sử dụng các thuật toán như Naive bayes, SVM (Support Vector Machine), KNN (K-nearest neighbor), cây quyết định với các thuật toán ID3 hay C4.5, mạng Neural; hồi quy - sử dụng các thuật toán như SVM. Mô hình đề xuất phân tích dữ liệu để đề xuất những quyết định phù hợp hoặc dự đoán các kết quả sẽ xảy ra.

**Phân lớp** là một phương pháp được sử dụng rất phổ biến trong khai phá dữ liệu. Nhiệm vụ chính của phân lớp là tổng quát hóa cấu trúc, lớp dữ liệu đã biết để áp dụng cho các đối tượng, mẫu dữ liệu mới. Nói cách khác, phân lớp có chức năng gắn các đối tượng vào tập chủ đề hay lớp mục tiêu dựa trên mức độ tương đồng. Mục đích của phân lớp là dự đoán lớp mục tiêu của mỗi trường hợp dữ liệu. Các kỹ thuật phân lớp phổ biến là Cây quyết định, láng giềng gần nhất, SVM hay Naive Bayes. Các thuật toán cây quyết định đáng chú ý như ID3, C4.5, CART, ... C4.5 là thuật toán sinh cây quyết định được phát triển bởi Ross Quinlan [3], cũng là phiên bản mở rộng của thuật toán ID3 ra đời trước đó.

**Thuật toán C4.5** là thuật toán phân lớp dữ liệu dựa trên cây quyết định; C4.5 sử dụng cơ chế lưu trữ dữ liệu thường trú trong bộ nhớ, chính đặc điểm này khiến C4.5 rất thích hợp với những cơ sở dữ liệu nhỏ, và cơ chế sắp xếp lại dữ liệu tại mỗi node trong quá trình phát triển cây quyết định. C4.5 còn chứa một kỹ thuật cho phép biểu diễn lại cây quyết định dưới dạng một danh sách sắp thứ tự các luật. Kỹ thuật này cho phép làm giảm bớt kích thước tập luật và đơn giản hóa các luật mà độ chính xác so với nhánh tương ứng cây quyết định là tương đương. Từ những lý do đó, thuật toán C4.5 rất phù hợp với bài toán xây dựng mô hình phân lớp trên dữ liệu của sinh viên đã tốt nghiệp khoa Công nghệ thông tin khi dữ liệu có quy mô không quá lớn.

### 4. Triển khai mô hình phân lớp

Nhóm nghiên cứu tiến hành thực hiện khai phá dữ liệu của các sinh viên đã tốt nghiệp thuộc khoa Công nghệ thông tin, trường Đại học Hàng hải Việt Nam. Việc triển khai mô hình phân lớp được thực hiện trên cơ sở dữ liệu sinh viên về mảng đào tạo, dữ liệu này được thu thập, tổng hợp sau mỗi kì học và được lưu trữ trên hệ thống máy chủ đào tạo. Một cách tổng quát, dữ liệu bao gồm tên sinh viên, ngành, chuyên ngành, lớp, điểm tích lũy (tương ứng với điểm trung bình học tập toàn khóa đối với sinh viên đã tốt nghiệp), xếp loại tốt nghiệp và thời gian tốt nghiệp. Phạm vi nghiên cứu: đối tượng được lựa chọn là các sinh viên đã tốt nghiệp của *ngành Công nghệ thông tin* trong vòng 04 khóa (*khóa 52, 53, 54 và 55*) với 03 chuyên ngành (*Công nghệ thông tin* với hai khóa 52-53, *Kỹ thuật phần mềm* và *Truyền thông - mạng máy tính* với cả bốn khóa). Ngoài ra, nhóm nghiên cứu tiến hành thực hiện một khảo sát tình hình việc làm của các sinh viên mới tốt nghiệp thuộc diện đối tượng nêu trên. Thông tin thu thập là cơ hội việc làm và lĩnh vực công việc, bao gồm *không có thông tin, chưa có việc làm, đang có việc làm và học tiếp*. Riêng đối với các đối tượng sinh viên *đang có việc làm*, khảo sát sẽ thu thập thêm thông tin những viên đó làm việc *đúng theo chuyên ngành học*, việc *có liên quan đến ngành học* hay *làm việc ngoài ngành học*. Mục đích chính của mô hình là phân tích kết quả thực hiện phân lớp tập dữ liệu theo hai thuộc tính *xếp loại tốt nghiệp* và *công việc* nhằm chỉ ra mối quan hệ giữa các thuộc tính này với nhau, cũng như mức độ ảnh hưởng của xếp loại tốt nghiệp tới cơ hội việc làm của sinh viên đã tốt nghiệp.

Nhóm nghiên cứu sử dụng công cụ WEKA [4] để hỗ trợ thực hiện các bước tiền xử lý và phân lớp dữ liệu. WEKA là một phần mềm học máy mã nguồn mở, được Đại học Waikato (New Zealand)

phát triển bằng Java, tập hợp các thuật toán học máy cho công việc khai phá dữ liệu, phần mềm bao gồm các công cụ cho việc chuẩn bị dữ liệu, phân lớp, hồi quy, phân cụm, khai phá tập luật và trực quan hóa.

### Tiền xử lý dữ liệu

Tập dữ liệu nghiên cứu về các đối tượng sinh viên nêu trên được trích xuất từ cơ sở dữ liệu sinh viên và lưu trữ thành tập *dlsv\_fit\_vmu.csv*. Tập dữ liệu thô bao gồm các bản ghi dữ liệu với nhiều trường thuộc tính, ví dụ mẫu về các bản ghi sinh viên được lấy ngẫu nhiên cùng các trường thể hiện ở Hình 1.

STT	Ma_sv	Ho_dem	Ten	Gioi_tinh	Ngay_sinh	Ten_lop	Trang_thai	Nam_hoc	TBCHT	TBCHT10	Xep_loai	So_tin_chi
1	56947	Đông Xuân	Việt	Nam	/1996	CNT55ĐH2	Tốt nghiệp	2018-2019	3.71	8.63	Xuất sắc	151
2	51298	Lê Ngọc	Hòa	Nam	/1995	CNT54ĐH2	Tốt nghiệp	2017-2018	3.69	8.62	Xuất sắc	150
3	56823	Nguyễn Mạnh	Hùng	Nam	/1996	CNT55ĐH1	Tốt nghiệp	2018-2019	3.66	8.51	Xuất sắc	154
4	51277	Trương Hàm	Yên	Nữ	/1995	CNT54ĐH1	Tốt nghiệp	2017-2018	3.63	8.51	Xuất sắc	150
5	43225	Đài Văn	Vinh	Nam	/1993	CNT52ĐH2	Tốt nghiệp	2015-2016	3.55	8.4	Giỏi	164
6	45317	Nguyễn Quốc	Manh	Nam	/1994	CNT53ĐH2	Tốt nghiệp	2016-2017	3.54	8.41	Giỏi	162
378	43122	Trần Viễn	Đông	Nam	/1992	CNT52ĐH1	Tốt nghiệp	2017-2018	2.00	5.89	Trung bình	165

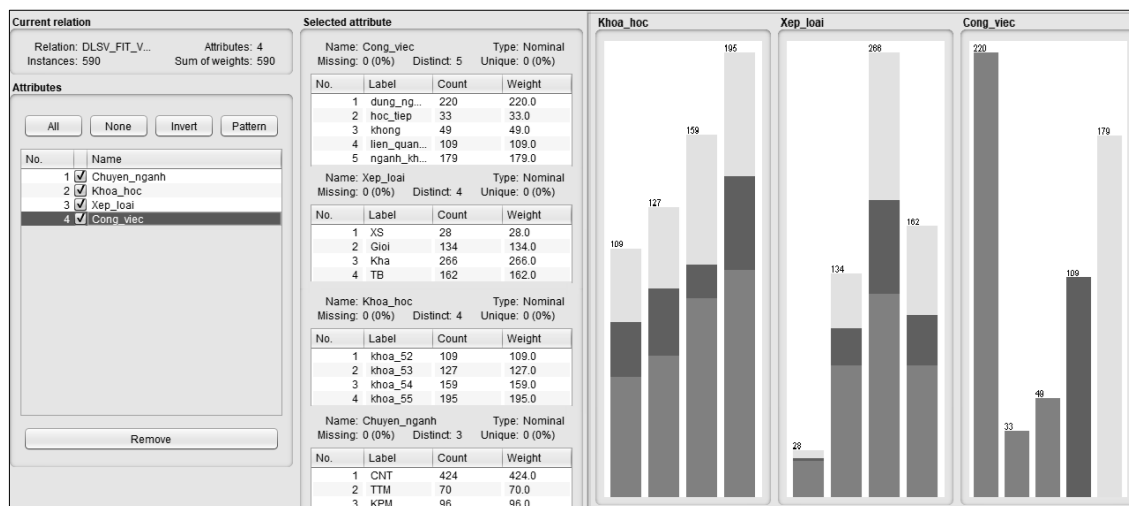
Hình 1. Một số bản ghi mẫu trong tập dữ liệu gốc *dlsv\_fit\_vmu.csv*

Tiếp theo, dữ liệu được tiền xử lý với công cụ *WEKA* thông qua các bước như nạp dữ liệu (chuyển sang định dạng *ARFF* file), lọc các thuộc tính (loại bỏ các thuộc tính dư thừa hoặc không phù hợp với mô hình) và giá trị của các thuộc tính. Các thuộc tính và giá trị thuộc tính dữ liệu sinh viên bao gồm: *Chuyen\_nganh* {*CNT*, *KPM*, *TTM*}; *Khoa\_hoc* {*khoa\_52*, *khoa\_53*, *khoa\_54*, *khoa\_55*}; *Xep\_loai* {*XS*, *Giỏi*, *Kha*, *TB*} và *Cong\_viec* {*dung\_nganh\_hoc*, *lien\_quan\_nganh\_hoc*, *nganh\_khac*, *hoc\_tiep*, *khong*}. Ví dụ mẫu về các bản ghi sinh viên sau khi tiền xử lý được lấy ngẫu nhiên và thể hiện trong Bảng 1.

Bảng 1. Một số bản ghi mẫu trong tập dữ liệu sau khi tiền xử lý

Chuyen_nganh	Khoa_hoc	Xep_loai	Cong_viec
CNT	khoa_52	XS	dung_nganh_hoc
CNT	khoa_53	Giỏi	dung_nganh_hoc
CNT	khoa_54	XS	hoc_tiep
KPM	khoa_54	Kha	lien_quan_nganh_hoc
KPM	khoa_55	TB	nganh_khac
TTM	khoa_54	Giỏi	lien_quah_nganh_hoc
TTM	khoa_55	TB	khong

Thống kê thông tin về các thuộc tính cùng các giá trị thuộc tính được cung cấp ở Hình 2.



Hình 2. Quá trình tiền xử lý dữ liệu

### Thực hiện phân lớp

Tiến hành phân lớp tập dữ liệu với thuật toán **C4.5** (cây **J48** của công cụ *WEKA*).

- Trường hợp phân lớp dữ liệu theo thuộc tính *Cong\_viec*, với chế độ sử dụng tập huấn luyện *train\_set* là toàn bộ tập dữ liệu *dataset*. Kết quả và thống kê của quá trình phân lớp được cung cấp trong Hình 3.

```

=== Run information ===
Scheme:      weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:    DLSV_FII_VMU
Instances:   590
Attributes:  4
              Chuyen_nganh
              Khoa_hoc
              Xep_loai
              Cong_viec

Test mode:   evaluate on training data

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Classifier model (full training set) ===
J48 pruned tree
-----
Xep_loai = XS: dung_nganh_hoc (28.0/8.0)
Xep_loai = Gioi: dung_nganh_hoc (134.0/69.0)
Xep_loai = Kha: dung_nganh_hoc (266.0/171.0)
Xep_loai = TB: nganh_khac (162.0/109.0)
Number of Leaves :    4
Size of the tree :    5
Time taken to build model: 0.01 seconds

=== Summary ===
Correctly Classified Instances      233      39.4915 %
Incorrectly Classified Instances    357      60.5085 %
Kappa statistic                     0.0636
Mean absolute error                 0.279
Root mean squared error             0.3735
Relative absolute error             96.1395 %
Root relative squared error         98.0932 %
Total Number of Instances          590

=== Confusion Matrix ===
      a  b  c  d  e  <-- classified as
180  0  0  0  40 | a = dung_nganh_hoc
 33  0  0  0  0 | b = hoc_tiep
 10  0  0  0  39 | c = khong
 79  0  0  0  30 | d = lien_quan_nganh_hoc
126  0  0  0  53 | e = nganh_khac

```

**Hình 3. Kết quả thực hiện phân lớp theo thuộc tính Cong\_viec (ở chế độ sử dụng toàn bộ dataset làm training\_set và test\_set)**

Các thông số về độ chính xác khi thực hiện phân lớp trong trường hợp này được cung cấp ở Hình 4 bao gồm TP/FP rate, F-measure, ROC area,...

```

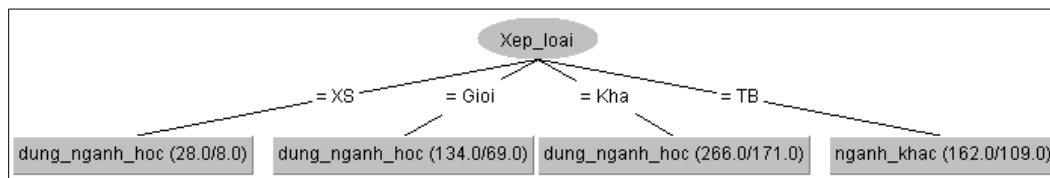
=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	FRC Area	Class
	0,818	0,670	0,421	0,818	0,556	0,160	0,624	0,469	dung_nganh_hoc
	0,000	0,000	?	0,000	?	?	0,690	0,090	hoc_tiep
	0,000	0,000	?	0,000	?	?	0,815	0,215	khong
	0,000	0,000	?	0,000	?	?	0,554	0,204	lien_quan_nganh_hoc
	0,296	0,265	0,327	0,296	0,311	0,032	0,548	0,326	nganh_khac
Weighted Avg.	0,395	0,330	?	0,395	?	?	0,607	0,334	

**Hình 4. Kết quả chi tiết về độ chính xác khi phân lớp theo thuộc tính Cong\_viec (ở chế độ full training set)**

Mô hình phân lớp được trực quan hóa thông qua cây quyết định (J48) được trình bày tại Hình 5, với 5 nút và 4 lá, thể hiện mối quan hệ giữa hai thuộc tính Xep\_loai và Cong\_viec.



**Hình 5. Cây quyết định (J48) thu được khi phân lớp theo thuộc tính Cong\_viec (ở chế độ full training set)**

- Nhằm mục đích so sánh kết quả và độ chính xác, tiếp tục thực hiện phân lớp dữ liệu theo thuộc tính Cong\_viec trong các trường hợp: chế độ *percentage split*: 66% (chia tập dữ liệu dataset thành hai phần với tỉ lệ 2/3 số bản ghi dùng làm tập huấn luyện training\_set và phần còn lại dùng làm tập kiểm tra test\_set) và chế độ *k-fold cross-validation*: 10 folds. Kết quả thu được và độ chính xác trong các trường hợp trên xấp xỉ với chế độ full training set, đồng thời các cây quyết định cũng có dạng tương tự.

- Cuối cùng, thực hiện phân lớp dữ liệu theo thuộc tính Xep\_loai với mục đích tìm hiểu mối liên hệ giữa thuộc tính Xep\_loai (xếp loại tốt nghiệp của sinh viên) với các thuộc tính còn lại. Kết quả về độ chính xác được thể hiện trong Hình 6. Cây quyết định thu được (Hình 7) trực quan hóa kết quả của việc phân lớp.

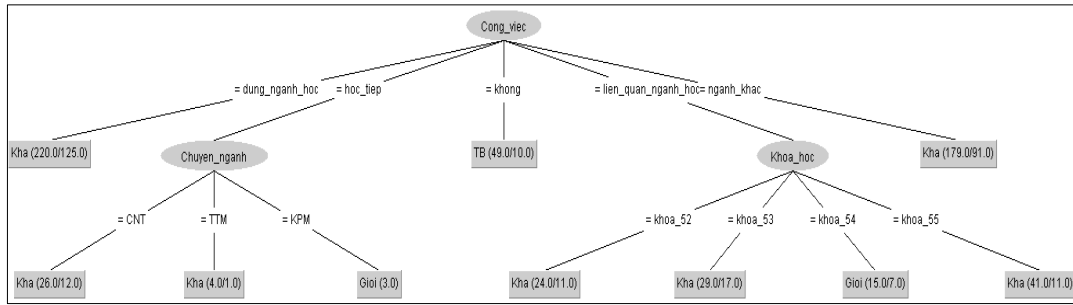
```

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	FRC Area	Class
	0,000	0,000	?	0,000	?	?	0,721	0,085	XS
	0,082	0,015	0,611	0,082	0,145	0,163	0,669	0,349	Gioi
	0,959	0,827	0,488	0,959	0,646	0,206	0,627	0,552	Kha
	0,241	0,023	0,796	0,241	0,370	0,352	0,712	0,495	TB
Weighted Avg.	0,517	0,383	?	0,517	?	?	0,665	0,468	

**Hình 6. Kết quả chi tiết về độ chính xác khi phân lớp theo thuộc tính Xep\_loai**



**Hình 7. Cây quyết định (J48) thu được khi phân lớp theo thuộc tính Xep\_loai**

## 5. Kết quả thực nghiệm và phân tích

Sau khi triển khai mô hình phân lớp dữ liệu của sinh viên tốt nghiệp theo thuộc tính *Cong\_viec* và *Xep\_loai* với thuật toán C4.5, các thông số thống kê thu được cũng như đánh giá độ chính xác của kỹ thuật phân lớp được cung cấp trong Hình 3, 4, 6. Kết quả thực nghiệm đã chỉ ra mối quan hệ giữa các thuộc tính *Chuyen\_nghanh*, *Khoa\_hoc*, *Xep\_loai* và *Cong\_viec*, cũng chính là sự liên quan giữa chuyên ngành học, khóa học và kết quả xếp loại tốt nghiệp đối với cơ hội nghề nghiệp của sinh viên. Quan trọng hơn, kết quả phân lớp được trực quan hóa thông qua các cây quyết định (Hình 5, 7) đã chỉ ra sự ảnh hưởng của xếp loại tốt nghiệp tới lựa chọn lĩnh vực công việc của các sinh viên đã ra trường, qua đó cung cấp dự đoán về cơ hội việc làm cho sinh viên sẽ ra trường trong tương lai. Kết quả cho thấy, đối với ngành Công nghệ thông tin, tỉ lệ sinh viên có việc làm sau khi tốt nghiệp cao, đặc biệt các sinh viên tốt nghiệp với xếp loại khá, giỏi có xu hướng làm việc đúng ngành học. Ngoài ra, số lượng sinh viên tốt nghiệp cùng với tỉ lệ sinh viên có kết quả tốt nghiệp cao (từ khá trở lên) tăng lên sau mỗi khóa học. Tuy nhiên, kết quả phân tích dữ liệu cũng ra có sự chênh lệch nhất định về chất lượng sinh viên tốt nghiệp của mỗi chuyên ngành, đồng thời số lượng sinh viên xếp loại tốt nghiệp xuất sắc và số sinh viên lựa chọn học tiếp vẫn còn ở mức rất thấp.

## 6. Kết luận

Trong bài viết này, nhóm nghiên cứu đã trình bày về chủ đề khai phá dữ liệu và các kỹ thuật phân lớp trong khai phá dữ liệu, áp dụng vào xây dựng mô hình phân lớp dữ liệu của sinh viên sử dụng thuật toán phân lớp C4.5. Kết quả thu được giúp phân tích mối quan hệ giữa điểm trung bình toàn khóa (xếp loại tốt nghiệp) và cơ hội việc làm của sinh viên đã tốt nghiệp, từ đó đưa ra các dự đoán hỗ trợ cho việc lựa chọn công việc của các sinh viên trong tương lai. Ở bài viết sau, nhóm nghiên cứu sẽ trình bày về phương pháp khai phá dữ liệu điểm các môn học của sinh viên nhằm tìm ra mối liên hệ và sự ảnh hưởng của kết quả các môn học tới điểm tích lũy toàn khóa học của sinh viên, sử dụng kỹ thuật khai phá luật kết hợp với thuật toán Apriori.

## TÀI LIỆU THAM KHẢO

- [1] Tariq O. Fadl Elsid, Mirghani. A. Eltahir, *Data Mining: Classification Techniques of Students' Database A Case Study of the Nile Valley University*, North Sudan, International Journal of Computer Trends and Technology Vol.16 No.5, 10/2014.
- [2] K. Sumathi, S. Kannan, K. Nagarajan, *Data Mining: Analysis of student database using Classification Techniques*, International Journal of Computer Applications Vol.141 No.8, 05/2016.
- [3] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.
- [4] J. R. Quinlan. *Improve Used of Continuous Attribute in C4.5*. In Journal of Artificial Intelligence Research 4 (1996) 77-90
- [5] Website: <https://www.cs.waikato.ac.nz/ml/WEKA/>.

Ngày nhận bài: 03/4/2019

Ngày nhận bản sửa: 23/4/2019

Ngày duyệt đăng: 03/5/2019