ĐÁNH GIÁ HỒ SƠ TUYỂN DỤNG BẰNG HỌC MÁY

Bùi Thanh Hùng[(1)]

(1) Trường Đại học Thủ Dầu Một

Ngày nhận bài 5/09/2020; Ngày gửi phản biện 10/09/2020; Chấp nhận đăng 20/10/2020

Liên hệ email: hungbt.cntt@tdmu.edu.vn

https://doi.org/10.37550/tdmu.VJS/2020.06.089

Tóm tắt

Trong cách mạng công nghiệp 4.0, việc áp dụng CNTT vào đời sống ngày càng thiết thực. Các công việc cũng cần có những xử lý của máy móc, trong đó có thể kể tới những bài toán phân tích và dự đoán kết quả của người tìm việc và người tuyển dụng. Các ứng viên tìm việc và nhà tuyển dụng cũng muốn có những thông tin và kết quả dự đoán chính xác nhằm có những đề xuất công việc phù hợp với bản thân mình. Nghiên cứu này được xây dựng dựa trên nhu cầu thực tế về việc ứng dụng công nghệ đánh giá hồ sơ tuyển dụng bằng học máy đáp ứng yêu cầu của người tìm việc và nhà tuyển dụng trong quá trình đánh giá hồ sơ tuyển dụng, đánh giá và đề xuất các công việc phù hợp với bộ hồ sơ. Chúng tôi đề xuất sử dụng 3 phương pháp học máy (Support Vector Machine - SVM, Decision Tree - DT, Random Forest - RF) để dự đoán hồ sơ tuyển dụng. Cơ sở đánh giá trên bộ dữ liệu của Trung tâm Giới thiệu việc làm tỉnh Bình Dương. Trên cơ sở phương pháp cho kết quả tốt nhất, chúng tôi xây dựng ứng dụng đánh giá hồ sơ tuyển dụng và trực quan hóa kết quả.

***Từ khóa**: đánh giá, hồ sơ tuyển dụng, học máy*

Abstract

EVALUATING RECRUITMENT PROFILE USING MACHINE LEARNING

In the era of industrial revolution 4.0, the application of IT has been playing a significant role. Analyzing and predicting the results of recruitment profile have gradually become the hot topic of interest to both researcher and business. By analyzing and predicting the recruitment profile, recruiters could evaluate candidate insights as well as predict which job is suitable for candidates. In this research, we propose evaluating recruitment profile using machine learning approach. We use Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) to evaluate recruitment profile. Our experiments on the dataset of the Binh Duong Job Center show the good results.

1. Giới thiệu

Sự phát triển mạnh mẽ của công nghệ thông tin làm cho máy tính trở thành phương tiện không thể thiếu được trong mọi lĩnh vực đời sống. Công nghệ thông tin càng phát triển thì con người càng có nhiều những phương pháp mới, công cụ mới để xử lý thông tin và nắm bắt được nhiều thông tin hơn. Công nghệ thông tin được ứng dụng trong mọi ngành nghề, mọi lĩnh vực sản xuất, kinh doanh, du lịch là một xu hướng

tất yếu. Kết quả của việc áp dụng công nghệ thông tin trong quản lý là việc hình thành các hệ thống thông tin quản lý nhằm phục vụ cho nhu cầu xử lý dữ liệu và cung cấp thông tin cho các chủ sở hữu hệ thống đó.

Trong kinh doanh mọi doanh nghiệp đều phải tiến hành tuyển dụng nhân sự. Công tác tuyển dụng nhân sự có một vai trò hết sức quan trọng, nó là tiền đề của bố trí, sử dụng và đào tạo phát triển. Tuyển dụng nhân sự được tiến hành thường xuyên bởi vì nhân sự của doanh nghiệp có thể biến động bất ngờ và ngẫu nhiên. Tuyển dụng nhân sự là một quy trình, được tiến hành qua nhiều bước, trong đó có một bước rất quan trọng đó là đánh giá ứng viên.

Đánh giá, lựa chọn ứng viên là quá trình so sánh nhiều ứng viên khác nhau với các tiêu chuẩn tuyển dụng để xác định ứng viên đáp ứng tốt nhất. So sánh các ứng viên là một việc khó, nhất là khi có rất nhiều các ứng viên. Vì vậy, trước khi tiến hành đánh giá ứng viên tổ chức cần xác định được quy trình và các tiêu chuẩn đánh giá cùng một phương pháp thống nhất để so sánh nhằm tìm ra ứng viên phù hợp nhất.

Có 2 phương pháp thường dùng để đánh giá, so sánh các ứng viên là xếp hạng và chấm điểm: (1) Phương pháp xếp hạng (ứng viên được xếp hạng theo các tiêu chuẩn tuyển dụng); (2) Phương pháp chấm điểm (để đánh giá, so sánh các ứng viên cần chấm điểm từng ứng viên theo các tiêu chuẩn xét tuyển; điểm cho mỗi tiêu chuẩn cần được quy định cụ thể). Phương pháp xếp hạng có nhược điểm là phải xác định được mức độ quan trọng của mỗi tiêu chuẩn trong đánh giá tổng thể. Việc xếp hạng không thể tiến hành được cho đến khi đã đánh giá xong tất cả các ứng viên. Nếu có nhiều ứng viên thì thật khó có thể nhớ chính xác thông tin của mỗi ứng viên.

Dù có các phương pháp đánh giá, so sánh các ứng viên, tuy nhiên không có phương pháp nào là hoàn hảo và tất cả chúng ta đều có thể cho điểm những ứng viên mà chúng ta thích cao hơn so với những ứng viên mà chúng ta không thích. Không phải dễ dàng có được sự đánh giá hoàn toàn khách quan, bởi vậy tổ chức phải linh hoạt áp dụng phương pháp đánh giá, so sánh ứng viên phù hợp với phương pháp tuyển dụng.

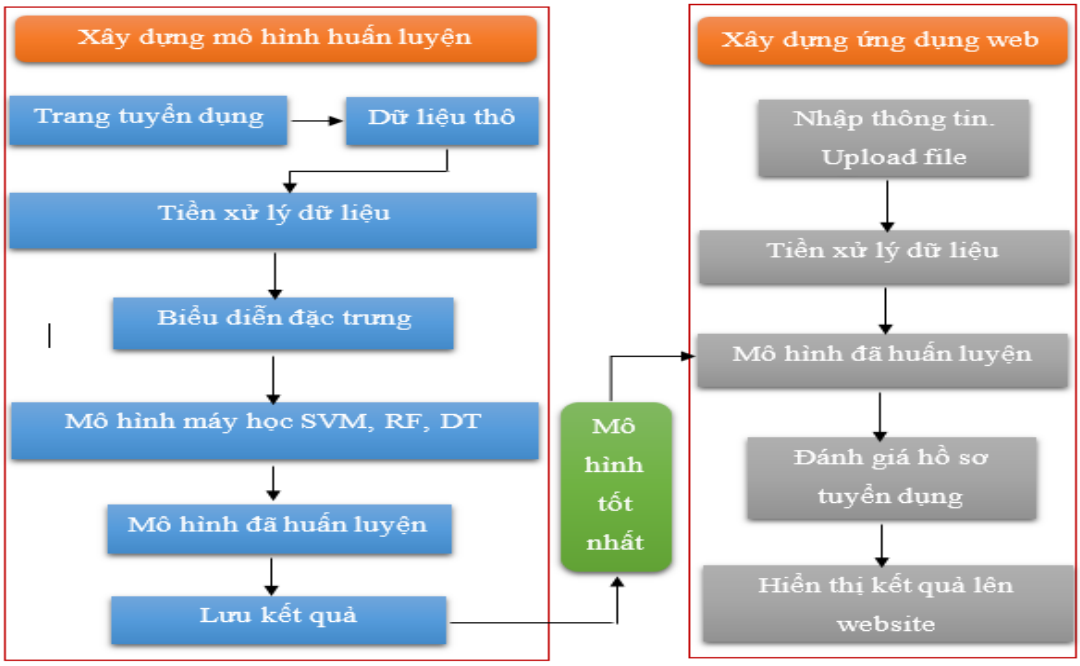
Đánh giá và lựa chọn gợi ý công việc phù hợp cho ứng viên là một quá trình gồm nhiều bước, mỗi bước trong quá trình là một phương pháp tuyển dụng. Số bước hay số phương pháp được sử dụng không cố định mà nó phụ thuộc vào mức độ phức tạp của công việc và tính chất của loại lao động cần tuyển dụng. Chính vì vậy cần có một ứng dụng đánh giá hồ sơ tuyển dụng một cách tự động để đề xuất công việc phù hợp với từng ứng viên.

Có nhiều cách tiếp cận cho vấn đề này, tuy nhiên đa số là tiếp cận theo hướng thủ công, sử dụng con người là chính. Một số nhà nghiên cứu đề xuất sử dụng học máy và áp dụng các giải pháp của xử lý ngôn ngữ tự nhiên để giải quyết bài toán này (FoDRA, 2016; Jayashree Rout, Sudhir Bagade, Pooja Yede, Nirmiti Patil, 2019) và đây là phương pháp chính giải quyết bài toán Đánh giá hồ sơ tuyển dụng. Trong nghiên cứu này chúng tôi cũng tiếp cận giải quyết bài toán bằng phương pháp học máy.

2. Mô hình đề xuất

2.1 Tổng quan về mô hình đề xuất

Mô hình tổng quát được trình bày trong Hình 1 với 2 phần chính đó là xây dựng mô hình đánh giá bằng phương pháp học máy và xây dựng ứng dụng demo chương trình cho người sử dụng và nhà tuyển dụng.



Hình 1. *Mô hình tổng quát*

Sau khi lấy dữ liệu từ trang tuyển dụng việc làm tỉnh Bình Dương và gán nhãn theo các loại công việc với các ứng viên. Các đặc trưng sẽ được biểu diễn thành dạng số và xây dựng thành các mô hình học máy dựa trên các thuật toán máy học Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF). Các dữ liệu sau khi đã được xử lý sẽ được chuyển đổi thành các vector số và đưa vào các mô hình máy học để huấn luyện phục vụ cho quá trình dự đoán. Ứng dụng được xây dựng trên các mô hình máy học đã được huấn luyện trước, dự đoán dựa trên các thông tin người dùng nhập vào và trả về kết quả hiển thị lên giao diện người dùng.

2.2. Đặc trưng

Dữ liệu huấn luyện bao gồm 2 dạng: dạng dữ liệu số và dạng dữ liệu chuỗi. Dữ liệu số bao gồm: tuổi, giới tính, số năm kinh nghiệm làm việc. Dữ liệu dạng chuỗi bao gồm: trình độ học vấn, ngành nghề trước đây, ngoại ngữ, tin học. Với mỗi loại dữ liệu được tiền xử lý, rút trích đặc trưng khác nhau để chuyển thành dữ liệu số và đưa vào mô hình huấn luyện. Các bước xử lý và chuyển hóa dữ liệu thành các vector đặc trưng được tiến hành như sau:

– Các dữ liệu dạng số: các giá trị này có các giá trị lớn nhỏ khác nhau tác động tới tính hiệu quả của nhiều thuật toán liên quan đến các vấn đề như thời gian thực hiện, quá trình hội tụ, độ chính xác của thuật toán. Do đó chúng ta cần một bước lý để chuẩn hóa các dữ liệu số thành các dữ liệu chuẩn Trong nghiên cứu này, chúng tôi sử dụng công thức sau để chuẩn hóa dữ liệu về dạng [0,1]:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (1)$$

– Các dữ liệu dạng chữ: Chúng tôi chuyển đổi dữ liệu văn bản thành vector, trước khi chuyển đổi, chúng tôi tiền xử lý dữ liệu bằng các bước:

Bước 1: Loại bỏ các dấu phẩy, dấu chấm, khoảng cách.

Bước 2: Tách từ tiếng Việt sử dụng thư viện Pyvi

Bước 3: Chuyển tất cả các từ về dạng chữ thường.

Sau khi đã tiền xử lý, chúng tôi chuyển đổi dữ liệu văn bản thành vector sử dụng phương pháp TF-IDF (Term Frequency – Inverse Document Frequency) (Stephen Robertson, 2004; Shahzad Qaiser,Ramsha Ali, 2018).

TF-IDF là 1 kĩ thuật sử dụng trong khai phá dữ liệu văn bản. Trọng số này được sử dụng để đánh giá tầm quan trọng của một từ trong một văn bản. Giá trị cao thể hiện độ quan trọng cao và nó phụ thuộc vào số lần từ xuất hiện trong văn bản nhưng bù lại bởi tần suất của từ đó trong tập dữ liệu. Công thức TF-IDF được trình bày như sau:

TF (Term Frequency): Tần suất xuất hiện của các từ

$$TF(t,d) = \frac{F(t,d)}{max(\{F(w,d) : w \in d\})} \quad (2)$$

Trong đó: TF(t, d): tần suất xuất hiện của từ t trong văn bản d; F(t, d): Số lần xuất hiện của từ t trong văn bản d; max({F(w, d) : w ∈ d}): Số lần xuất hiện của từ có số lần xuất hiện nhiều nhất trong văn bản d; IDF: Giúp đánh giá tầm quan trọng của một từ. Khi tính toán TF, tất cả các từ được coi như có độ quan trọng bằng nhau.

$$IDF(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \quad (3)$$

Trong đó: IDF(t, D): giá trị idf của từ t trong tập văn bản; |D|: Tổng số văn bản trong tập D; {d ∈ D : t ∈ d}|: thể hiện số văn bản trong tập D có chứa từ t

Công thức tính TF-IDF dựa trên TF và IDF như sau:

$$\mathrm{TF-IDF}(t,d,D) = \mathrm{TF}(t,d) * \mathrm{IDF}(t,D) \quad (4)$$

Chúng tôi sử dụng kỹ thuật TF-IDF để biểu diễn các cột thông tin dữ liệu văn bản của các cột thông tin của ứng viên. Tất cả các thông tin ở mỗi cột sẽ được thu thập lại tạo một tập từ điển các từ vựng có trong cột đó. Dựa vào tập từ điển này, mỗi giá trị thông tin của ứng viên được được biểu diễn bằng các vector dựa trên tập từ điển, sau đó công thức TF-IDF được tính trên từng vector và đưa ra vector đại diện cho từng thông tin của ứng viên.

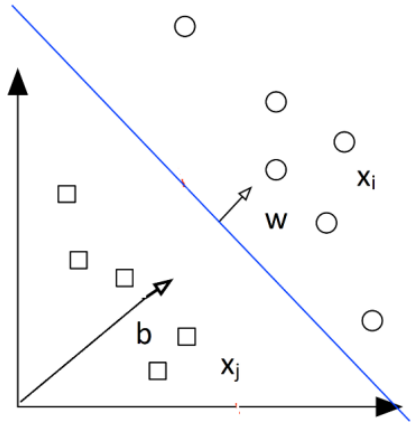
2.3. Huấn luyện

Chúng tôi sử dụng các phương pháp học máy véc tơ hỗ trợ – Support Vector Machine (SVM), Cây quyết định – Decision Tree (DT) và Rừng ngẫu nhiên – Random Forest (RF) để huấn luyện mô hình.

2.3.1. SVM

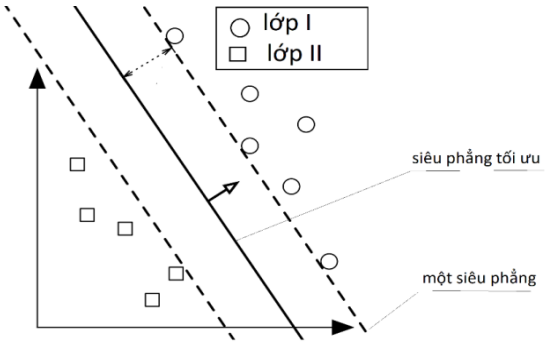
Phương pháp học máy véctơ hỗ trợ SVM ra đời từ lý thuyết học thống kê do Vapnik và Chervonekis xây dựng năm 1995 (Tom Mitchell, 1997; Jiawei Han, Micheline Kamber, 2006) và có nhiều tiềm năng phát triển về mặt lý thuyết cũng như ứng dụng trong thực tế. Phương pháp SVM có khả năng phân loại khá tốt đối với bài toán phân lớp cũng như trong nhiều ứng dụng thực tế.

Support Vector Machines (SVM) là kỹ thuật mới đối với việc phân lớp dữ liệu, là phương pháp học sử dụng không gian giả thuyết các hàm tuyến tính trên không gian đặc trưng nhiều chiều, dựa trên lý thuyết tối ưu và lý thuyết thống kê. Trong kỹ thuật SVM không gian dữ liệu nhập ban đầu sẽ được ánh xạ vào không gian đặc trưng và trong không gian đặc trưng này mặt siêu phẳng phân chia tối ưu sẽ được xác định. Hình 2 biểu diễn Phân tách theo siêu phẳng(w,b) trong không gian 2 chiều.



Hình 2. *Phân tách theo siêu phẳng(w,b) trong không gian 2 chiều*

Siêu phẳng có khoảng cách với dữ liệu gần nhất là lớn nhất (tức có biên lớn nhất) được gọi là siêu phẳng tối ưu. Hình 3 biểu diễn 1 siêu phẳng tối ưu.



Hình 3. *Siêu phẳng tối ưu*

Mục đích đặt ra ở đây là tìm được một ngưỡng (w,b) phân chia tập mẫu vào các lớp có nhãn 1 (lớp I) và -1 (lớp II) nêu ở trên với khoảng cách là lớn nhất. Như vậy, ý tưởng của SVM là đi tìm một mặt siêu phẳng để phân lớp dữ liệu. Các mặt phẳng được biểu diễn dưới dạng: $w^T x + b$ (5)

Khoảng cách của một siêu phẳng được tính theo công thức:

$$margin = \min_{n} \frac{y_n(w^T x_n + b)}{\|w\|_2} \quad (6)$$

Để xác định được một khoảng cách lớn nhất ta đi tìm w và b:

$$(w,\ b) = \arg\max_{w,b} \left\{ \min_{n} \frac{y_n(w^T x_n + b)}{\|w\|_2} \right\} \quad (7)$$

Hay : $(w, b) = \arg\min_{w,b} \frac{1}{2}\|w\|_2^2$ (8)

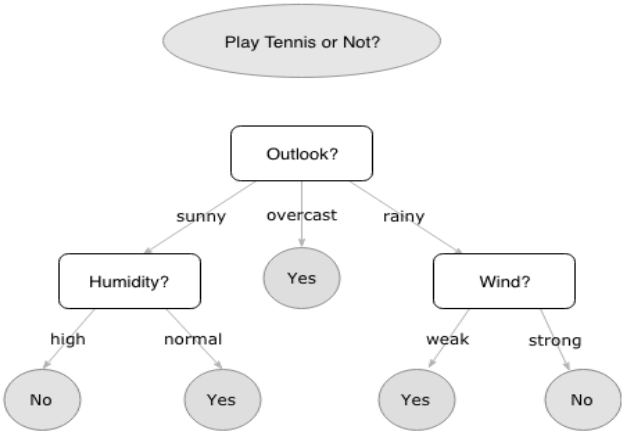
Như vậy vấn đề đặt ra ở đây là tìm w và b theo các công thức:

$$\lambda = arg\ \min_{\lambda} \frac{1}{2}\lambda^T K\lambda + p^T \quad (9)$$

Với : $\lambda \leq h$ và $A\lambda = b$; λ là hệ số cần tìm; K là một ma trận vuông; $G, A \in R^{(m \times n);}$ $h, b \in R^m$; $p \in R^n$

2.3.2. Cây quyết định

Cây quyết định là một kiểu mô hình dự báo (predictive model), nghĩa là một ánh xạ từ các quan sát về một sự vật/hiện tượng tới các kết luận về giá trị mục tiêu của sự vật/hiện tượng [10-11]. Mỗi một nút trong (internal node) tương ứng với một biến; đường nối giữa nó với nút con của nó thể hiện một giá trị cụ thể cho biến đó. Mỗi nút lá đại diện cho giá trị dự đoán của biến mục tiêu, cho trước các giá trị của các biến được biểu diễn bởi đường đi từ nút gốc tới nút lá đó. Kỹ thuật học máy dùng trong cây quyết định được gọi là học bằng cây quyết định, hay chỉ gọi với cái tên ngắn gọn là cây quyết định.

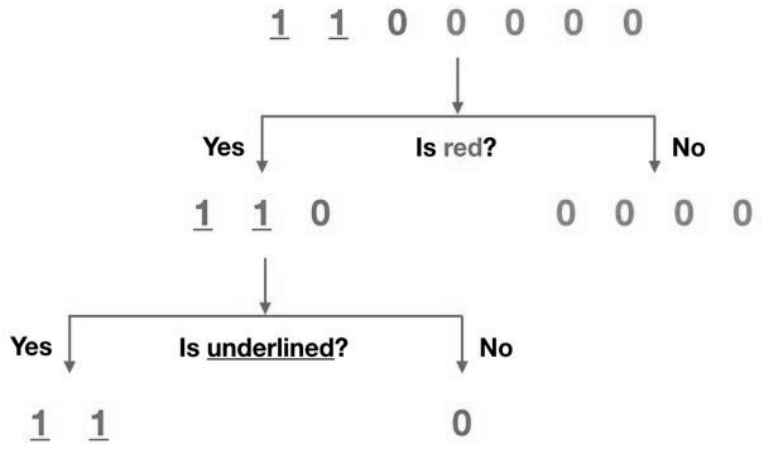


Hình 4. *Mô hình cây quyết định*

Cây quyết định là một mô hình máy học có giám sát, có thể được áp dụng vào cả hai bài toán phân loại và hồi quy. Việc xây dựng một cây quyết định trên dữ liệu huấn luyện cho trước là việc đi xác định các câu hỏi và thứ tự của chúng. Một điểm đáng lưu ý của Decision Tree là nó có thể làm việc với các đặc trưng dạng Categorical, thường là rời rạc và không có thứ tự.

2.3.3. Rừng ngẫu nhiên

Random Forests (RF) là một thuật toán học giám sát (supervised). "Ensemble" ở đây có nghĩa là tập hợp tất cả "weak learners" và giúp nó làm việc cùng nhau để tạo ra một dự báo có độ tin cậy cao [12]. Trong trường hợp này, những "weak learners" là tất cả các Decision Trees ngẫu nhiên được kết hợp để tạo thành dự đoán có độ tin cậy cao – Random Forest là một trong những thuật toán machine learning phổ biến nhất và mạnh nhất. Nó là một loại thuật toán machine learning được gọi là Bootstrap Aggregation hoặc Bagging. Hình 5 mô tả về mô hình rừng ngẫu nhiên.



Hình 5. *Mô hình rừng ngẫu nhiên*

3. Thực nghiệm

3.1 Dữ liệu

Dữ liệu trong đề tài này được thu thập trực tiếp từ trang tuyển dụng Việc làm Bình Dương của Trung tâm Giới thiệu việc làm tỉnh Bình Dương. Bộ dữ liệu thô này bao gồm 1967 mẫu dữ liệu hồ sơ mà người dùng đưa cho thông tin. Bộ dữ liệu được lưu dưới dạng câu trúc định dạng Excel bao gồm 13 cột thông tin khác nhau như: Họ Tên, Ngày Sinh, Giới Tính, Số CMND, Điện Thoại, Địa Chỉ Số, Lần Đăng Ký, Vị Trí Công Việc, Nơi Làm Việc, Năm Kinh Nghiệm, Trình Độ, Ngành, Ngoại Ngữ, Tin Học.

Xử lý dữ liệu

Chúng tôi tiến hành tiền xử lý dữ liệu bằng cách loại bỏ các thông tin liên quan đến thông tin cá nhân như: Họ tên, Số CMND, Điện Thoại, Địa Chỉ, Số lần đăng ký và các thông tin các cột còn lại. Cột Ngày sinh sẽ chuyển thành số tuổi tính đến thời điểm hiện tại. Tuy nhiên trong dữ liệu vẫn tồn tại nhiều thông tin lặp lại giữa các ứng viên, do đó để tránh nhiễu trong quá trình làm dữ liệu, chúng tôi loại bỏ các dữ liệu trùng chỉ giữ lại một dựa trên cột CMND mà các ứng viên nhập vào vì đây là giá trị định dang giữa các ứng viên với nhau. Kết quả sau khi lọc trùng trong bộ dữ liệu, chúng tôi thu được là 1.516 mẫu dữ liệu gán nhãn. Sau khi loại bỏ các thông tin không hữu ích, tiến hành gán loại ngành nghề phù

hợp cho từng mẫu hồ sơ. Bảng chi tiết ngành nghề được người thực hiện tham khảo trực tiếp từ trang websites của Trung tâm giới thiệu việc làm Tỉnh Bình Dương. Tiếp đó chúng tôi tiến hành gán nhãn dữ liệu các giá trị ứng viên nhập vào trong bộ dữ liệu. Dựa vào các cột thông tin mà các ứng viên nhập vào: ngành nghề, vị trí công việc mà họ mong muốn để tiến hành phân loại ra các ngành nghề như trên. Trong quá trình gán nhãn cũng có một số lỗi chính tả, không đúng cú pháp và các thông tin không hợp lệ cũng được tiền xử lý lại cho đúng với thông tin các cột. Chi tiết các thông tin cột dữ liệu mà các ứng viên cung cấp việc làm gồm: Tuổi, Giới tính, Năm kinh nghiệm, Trình độ, Ngoại ngữ, Tin học. Dữ liệu được chia thành 2 phần Train và Test theo tỉ lệ 8:2

3.2 Huấn luyện mô hình

Từ dữ liệu thô sau khi được tiền xử lý dữ liệu chúng tôi sử dụng kỹ thuật MinMaxScaling để đưa các giá trị số về dạng vector có dạng [0,1]. Đối với các dữ liệu văn bản, người thực hiện sử dụng kỹ thuật TF-IDF để đưa các giá trị văn bản về dạng vector biểu diễn. Sau đó nối các vector của các cột này lại với nhau để làm vector đại diện cho mỗi dòng dữ liệu, nhãn công việc cũng được chuyển thành dạng số tương ứng, người thực hiện đưa 2 giá trị này vào mô hình máy học thuật toán SVM, RF, DT để huấn luyện và đánh giá mô hình.

3.3 Kết quả thực nghiệm

Bộ dữ liệu sau khi được tiền xử lý chuyển thành vector và đưa vào huấn luyện bằng ba phương pháp học máy là Support Vector Machine (SVM), Rừng ngẫu nghiên (Random Forrest) và Cây quyết định (Decision Tree). Chúng tôi sử dụng ngôn ngữ lập trình Python, thư viện pyvi của Trần Việt Trung (2016) để tách từ, thư viện học máy Sklearn cùng với Numpy và Scipy, thiết kế giao diện ứng dụng bằng HTML, Javascrip, CSS và Boostrap. Kết quả được đánh giá trên ba độ đo: độ chính xác, độ bao phủ và độ đo F1 score. Các độ đo này được tính theo các công thức dưới đây.

$$\text{Precision} = TP/(TP + FP) \qquad (10)$$

Độ chính xác là khả năng của thuật toán phân loại không gán cho cho mẫu positive giá trị negative. Đối với mỗi class, nó được định nghĩa là tỷ lệ True Positive so với tổng True Positive và False Positive.

$$\text{Recall} = TP/(TP+FN) \qquad (11)$$

Độ phủ là khả năng một thuật toán phân lớp tìm ra các mẫu positive. Đối với mỗi class nó định nghĩa là tỷ lệ True Positive so với tỷ lệ True Positive với False Negative.

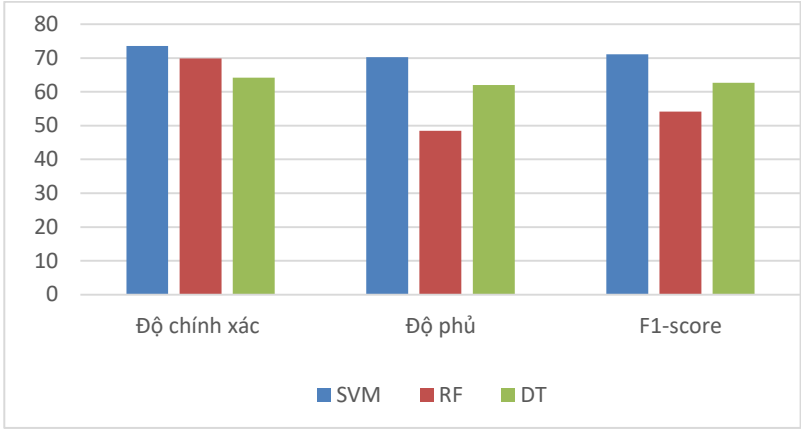
F1 score là chỉ số trung hòa giữa giá trị Precision và Recall.

$$\text{F1} - \text{score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (12)$$

Kết quả được trình bày trong Bảng 1. Hình 6 biểu diễn so sánh kết quả của 3 phương pháp SVM, RF và DT.

***Bảng 1**. Kết quả của 3 phương pháp SVM, RF, DT*

Phương pháp	Độ chính xác	Độ phủ	F1-score
SVM	**73.64**	**70.29**	**71.10**
RF	69.87	48.51	54.21
DT	64.21	62.05	62.73



***Hình 6.** So sánh kết quả của 3 phương pháp SVM, RF, DT*

***Hình 7**. Giao diện người dùng*

Dựa vào kết quả trong Bảng 1 ta thấy rằng phương pháp SVM đạt kết quả tốt nhất với độ chính xác là 73.64%, độ phủ là 79.29% và chỉ số F1-score là 71,10%. Kết quả cao hơn nhiều so với hai phương pháp còn lại là Random Forest hay Decision Tree. Do đó phương pháp SVM sẽ được lưu lại phục vụ cho ứng dụng minh họa trực quan hóa kết quả.

Ứng dụng được trực quan hóa kết quả hiển thị trên website gồm các chức năng: Hướng dẫn sử dụng, Ứng dụng đánh giá, Phân tích dữ liệu, Đánh giá kết quả nghiên cứu. Người dùng sẽ nhập trực tiếp các thông tin của ứng viên: Tuổi, Bằng cấp, Kinh Nghiệm, Ngoại Ngữ, Tin học, Giới Tính. Ứng dụng sẽ lấy các thông tin này tiền xử lý và đưa qua mô hình SVM để dự đoán và đưa ra ngành nghề phù hợp nhất và 4 ngành nghề phù hợp khác sắp xếp theo thứ tự từ cao xuống thấp gợi ý cho người dùng. Hình 7 biểu diễn giao diện người dùng và Hình 8 biểu diễn kết quả gợi ý.

Từ kết quả huấn luyện, có thể nhận thấy rằng kết quả dự đoán giữa nhân viên kinh doanh (NVKD) và nhân viên văn phòng (NVVP) đạt kết quả thấp nhất. Điều này một phần vì nhãn giữa hai loại này có cùng đặc trưng tương đối gần giống nhau, do đó tỷ lệ nhầm lẫn giữa hai nhãn này cao. Các ngành nghề khác như Bảo vệ (BV) hay nhân viên phiên dịch (NVPD) lại có kết quả rất cao trong bộ dữ liệu vì các nhãn này thường có các đặc trưng khác biệt với các nhãn khác.

Hình 8. *Kết quả đánh giá và gợi ý*

4. Kết luận

Nghiên cứu này trình bày một phương pháp đánh giá hồ sơ tuyển dụng bằng học máy. Dựa trên dữ liệu đầu vào được chuẩn hóa và chuyển đổi thành vector đặc trưng TF-IDF và huấn luyện bằng các mô hình học máy: SVM, Decision Tree và Random Forest. Qua thực nghiệm cho thấy, phương pháp học máy SVM cho kết quả tốt nhất. Chúng tôi cũng đã xây dựng ứng dụng đánh giá hồ sơ tuyển dụng trực tuyến và bước đầu khảo sát ghi nhận phản hồi của người sử dụng. Trong thời gian tới, chúng tôi sẽ tìm cách nghiên cứu xử lý các dữ liệu thu thập và thử nghiệm trên các mô hình khác để tìm được giải pháp tối ưu nhất cho việc đánh giá hồ sơ tuyển dụng.

TÀI LIỆU KHAM KHẢO

[1] FoDRA – Nikolaos D. Almalis George A. Tsihrintzis, Aggeliki D Strati (2016). "A New Content-Based Job Recommendation Algorithm for Job Seeking and Recruiting".

[2] Data,Vishnu M Menon Computer Rahul Nath H A (2016). "A Novel Approach to Evaluate and Rank Candidates in A Recruitment Process by Estimating Emotional Intelligence through Social Media".

[3] Manasi Ombhase, Prajakta Gogate, Tejas Patil (2017). Automated Personality Classification Using Data Mining Technoques. DOI: 10.13140/RG.2.2.35949.59363.

[4] Vivian Lai, Kyong Jin Shim, Richard J. Oentaryo, Philips K. Prasetyo, Casey Vu Ee-Peng Lim, David Lo (2016). "Career Mapper: An Automated Resume Evaluation Tool".

[5] Jayashree Rout, Sudhir Bagade, Pooja Yede, Nirmiti Patil (2019). Personality Evaluation and CV Analysis using Machine Learning Algorithm. *International Journal of Computer Sciences and Engineering*, Vol. 7, Issue 5.

[6] Stephen Robertson (2004). "Understanding inverse document frequency: on theoretical arguments for IDF". *Journal of Documentation*, Vol. 60, Issue 5.

[7] Shahzad Qaiser,Ramsha Ali (2018). "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents". *International Journal of Computer Applications*, Vol. 181, Isuue 1.

[8] Tom M. Mitchell (1997). *Machine Learning.* McGraw Hill, Inc.

[9] Jiawei Han, Micheline Kamber (2006). *Data Mining: Concepts and Techniques*. Second Edition. Morgan Kaufmann Publishers.

[10] Leo Breiman, Jerome Friedman, Charles J. Stone & R.A. Olshen (1984). *Classification and Regression Trees*. Taylor & Francis.

[11] Mihaela van der Schaar (2017). *Classification and regression trees.* Department of Engineering Science University of Oxford.

[12] Breiman, L. (2001). "Random forests". *Machine Learning*, Vol. 45, N^{o} 1.