**TRƯỜNG ĐẠI HỌC THỦY LỢI**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**BÁO CÁO MÔN HỌC MÁY**

**Ứng dụng phương pháp ID3 và PERCEPTRON dự đoán cơ hội việc làm cho sinh viên**

**GIẢNG VIÊN MÔN HỌC :** NGUYỄN THỊ KIM NGÂN

**NHÓM SINH VIÊN THỰC HIỆN (NHÓM 13):**

Phạm Minh Tiến - 1851061636

Trần Đức Huy - 1851060586

Phạm Hoài Nam - 1851061318

*Hà Nội, 09/2021*

### Mô tả bài toán

1. Tên đề tài: Sử dụng phương pháp ID3 và PERCEPTRON dự đoán cơ hội việc làm
2. Thuật toán sử dụng

* ID3: Cây quyết định (Decision Tree) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật. Các thuộc tính của đối tượng có thể thuộc các kiểu dữ liệu khác nhau như Nhị phân (Binary) , Định danh (Nominal), Thứ tự (Ordinal), Số lượng (Quantitative) trong khi đó thuộc tính phân lớp phải có kiểu dữ liệu là Binary hoặc Ordinal.
* PERCEPTRON: là một thuật toán Classification cho trường hợp đơn giản nhất: chỉ có hai class (lớp) (*bài toán với chỉ hai class được gọi là binary classification*) và cũng chỉ hoạt động được trong một trường hợp rất cụ thể.

1. Tóm tắt công việc:

* Thu thập dữ liệu của 627 công dân .
* Sử dụng 2 phương pháp ID3 và PERCEPTRON dự đoán cơ hội việc làm (PlacedOrNot) của mẫu thử .

### Mô tả dữ liệu của bài toán:

1. Vector dữ liệu gồm: 627 vector dữ liệu.
2. Chiều thông tin dữ liệu : Age , Gender , Stream , Internships , CGPA , Hostel , HistoryOfBacklogs , PlacedOrNot .
3. Đầu vào của bài toán gồm : Age , Gender , Stream , Internships ,CGPA , Hostel , HistoryOfBacklogs .
4. Thông tin đầu ra cần dự đoán : PlacedOrNot .

### III. Chia dữ liệu thành 2 phần

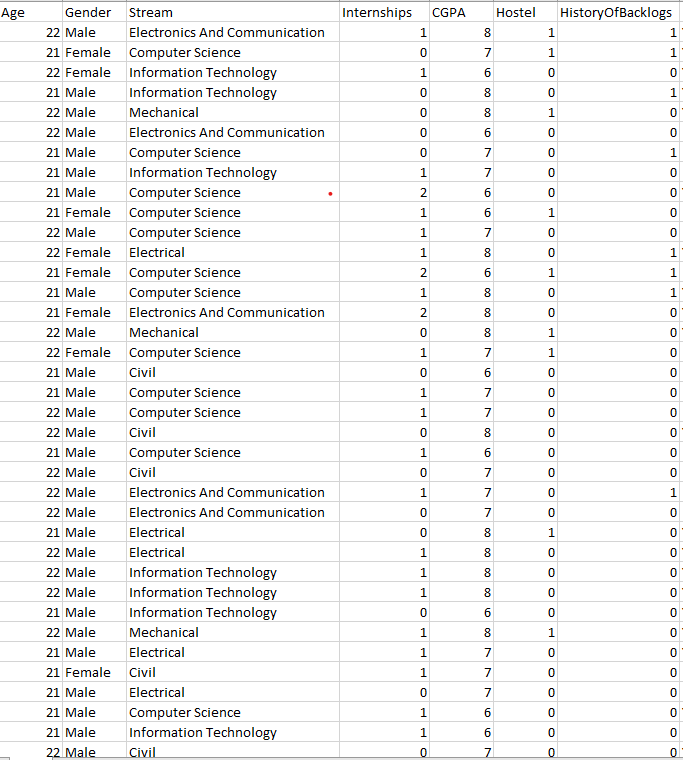
Dữ liệu được làm 2 phần bao gồm training và test  
 -Training : gồm 70% dữ liệu đầu tiên

-Test : gồm 30% dữ liệu còn lại

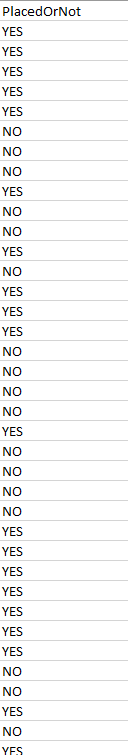
### IV . Mô tả dữ liệu huấn luyện

Trích dẫn 1 phần trong dữ liệu huấn luyện :

* Dữ liệu huấn luyện(X)



* Vecto đầu ra (Y):



### VI. So sánh kết quả dự đoán và kết quả thực

### 1,ID3(entropy)

ID3 cho ta tỉ lệ dự đoán như sau :

Số dự đoán đúng 157 trên tổng 189

Tỷ lệ đúng đạt : 83 %

### 2,Perceptron

Perceptron cho ta tỉ lệ dự đoán như sau :

Số dự đoán đúng 143 trên tổng 189

Tỷ lệ đúng đạt : 76 %

KL : Dựa vào tỉ lệ dự đoán ta có thể thấy được kết quả khi sử dụng thuật toán ID3 có tỉ lệ chính xác cao hơn perceptron

### VII. Kết Luận

## Ứng dụng

* ID3 (cây quyết định)
  + Mô hình sinh ra các quy tắc dễ hiểu cho người đọc, tạo ra bộ luật với mỗi nhánh lá là một luật của cây.
  + Dữ liệu đầu vào có thể là là dữ liệu missing, không cần chuẩn hóa hoặc tạo biến giả
  + Có thể làm việc với cả dữ liệu số và dữ liệu phân loại
  + Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê
  + Có khả năng là việc với dữ liệu lớn
* Perceptron là một thuật toán Classification cho trường hợp đơn giản nhất: chỉ có hai class (lớp) (*bài toán với chỉ hai class được gọi là binary classification*) và cũng chỉ hoạt động được trong một trường hợp rất cụ thể. Tuy nhiên, nó là nền tảng cho một mảng lớn quan trọng của Machine Learning là Neural Networks và sau này là Deep Learning.

## Hạn chế của phương pháp ID3 và Perceptron

* + **ID3**
* Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu của bạn. Thạm chí, với một sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc mô hình cây quyết định có thể thay đổi hoàn toàn.
* Cây quyết định hay gặp vấn đề [overfitting](https://trituenhantao.io/tu-dien-thuat-ngu/overfitting/)
* **Perceptron** 
  + Giá trị đầu ra của mạng perceptron chỉ có thể là 2 giá trị 0 hoặc 1.
  + Chỉ có thể phân loại tuyến tính. Nếu là đường thẳng hay mặt phẳng ta có thể vẽ tách rời các vector đầu vào thành các loại chính xác, các vector vào là độc lập tuyến tính. Nếu không phải là độc lập tuyến tính, việc học sẽ không bao giờ đạt tới mức các vector được phân loại chính xác. Tuy nhiên, điều đó cũng chứng minh được nếu các vector độc lập tuyến tính, việc huấn luyện perceptron sẽ luôn tìm được đáp án trong thời gian hữu hạn.