**TRƯỜNG ĐẠI HỌC THỦY LỢI**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**BÁO CÁO MÔN HỌC MÁY**

**Ứng dụng thuật toán Support-Vector-Machine và Cross-Validation để dự đoán cơ hội việc làm cho sinh viên năm cuối**

**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*

## Phần 1: Tổng quan

#### **1. Giới thiệu về học máy**

* Lịch sử và vai trò của machine learning
  + Lịch sử: **Học máy** ([tiếng Anh](https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Anh): *machine learning*) là một lĩnh vực của [trí tuệ nhân tạo](https://vi.wikipedia.org/wiki/Tr%C3%AD_tu%E1%BB%87_nh%C3%A2n_t%E1%BA%A1o) liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể. Ví dụ như các máy có thể "học" cách phân loại [thư điện tử](https://vi.wikipedia.org/wiki/Th%C6%B0_%C4%91i%E1%BB%87n_t%E1%BB%AD) xem có phải [thư rác (spam)](https://vi.wikipedia.org/wiki/Th%C6%B0_r%C3%A1c) hay không và tự động xếp thư vào thư mục tương ứng. Lịch sử của Học máy là một quá trình dài dựa trên các thuật toán từ cả trăm năm trước, cũng như những phát triển về công nghệ máy tính trong 7 thập kỷ qua.
    - Bước đột phá lớn – công trình của Thomas Bayes vào thế kỷ thứ 18 mà sau đó dẫn tới Định lý Bayes (1812) của Pierre-Simon Laplace
    - 1950 – Alan Turing tạo ra phép thử Turing (Turing Test) để xác định xem một máy tính có trí thông minh thực sự hay không.
    - 1952 – Arthur Samuel đã viết chương trình máy học đầu tiên. Chương trình này là trò cờ đam, và máy tính IBM càng chơi càng tiến bộ.
    - 1957 – Frank Rosenblatt thiết kế hệ thần kinh (neural network) đầu tiên cho máy tính (gọi là “perceptron”) mô phỏng các quá trình tư duy của não người.
    - 1967 – Thuật toán “điểm lân cận gần nhất” (nearest neighbor).
    - 1981 – Gerald Dejong giới thiệu khái niệm Học qua giải thích (Explanation Based Learning – EBL)
    - 1990s – Các nghiên cứu về machine learning chuyển từ cách tiếp cận hướng-kiến-​​thức sang cách tiếp cận hướng-dữ-liệu.
    - 1997 – Máy tính Deep Blue của IBM đánh bại nhà vô địch cờ vua thế giới.
    - 2006 – Geoffrey Hinton tạo ra thuật ngữ “học sâu” (deep learning).
    - 2010 – Máy Microsoft Kinect có thể theo dõi 20 tính năng của con người ở tốc độ 30 lần mỗi giây
    - 2011 – Máy IBM Watson đánh bại các đối thủ con người tại Jeopardy.
    - 2011 – Google Brain được phát triển, và hệ thần kinh sâu (deep neural network).
    - 2012 – X Lab của Google phát triển một thuật toán machine learning có khả năng tự duyệt video YouTube để xác định các video có chứa hình ảnh mèo.
    - 2014 – Facebook phát triển DeepFace, một thuật toán phần mềm có thể nhận diện hoặc xác minh các cá nhân trong các bức ảnh ở mức độ ngang với con người.
    - 2015 – Microsoft tạo ra bộ công cụ máy học được phân bổ (Distributed Machine Learning Toolkit), cho phép phân bổ hiệu quả machine learning trên nhiều máy tính
    - 2015 – Hơn 3.000 nhà nghiên cứu về AI và Robotics, được Stephen Hawking, Elon Musk và Steve Wozniak (và nhiều người khác) hậu thuẫn, đã ký một bức thư ngỏ cảnh báo nguy cơ của vũ khí tự hành có thể lựa chọn và nhắm vào các mục tiêu mà không cần sự can thiệp của con người.
    - 2016 – Thuật toán trí tuệ nhân tạo (artificial intelligence algorithm) của Google đã đánh bại một cầu thủ chuyên nghiệp trong trò cờ Go của Trung Quốc, vốn được coi là trò cờ phức tạp nhất thế giới và khó hơn cờ vua gấp nhiều lần. Thuật toán AlphaGo do Google DeepMind phát triển đã giành chiến thắng năm trên năm trận đấu Go.
  + Vai trò:
    - Xử lý ảnh: Bài toán xử lý ảnh(Image Processing) giải quyết các vấn đề phân tích thông tin từ hình ảnh hay thực hiện một số phép biến đổi
    - Phân tích văn bản: Phân tích văn bản(Text analysis) là công việc trích xuất hoặc phân loại thông tin từ văn bản. Các văn bản ở đây có thể là các facebook posts, emails, các đoạn chats, tài liệu,…
    - Khai phá dữ liệu: Khai phá dữ liệu(Data mining) là quá trình khám phá ra các thông tin có giá trị hoặc đưa ra các dự đoán từ dữ liệu. Định nghĩa này có vẻ bao quát, nhưng bạn hãy nghĩ về việc tìm kiếm thông tin hữu ích từ một bảng dữ liệu rất lớn. Mỗi bản ghi sẽ là một đối tượng cần phải học, và mỗi cột là một đặc trưng. Chúng ta có thể dự đoán giá trị của một cột của bản ghi mới dựa trên các bản ghi đã học. Hoặc là phân nhóm các bản ghi của bản
    - Trò chơi điện tử & Robot: Trò chơi điện tử(Video games) và robot(Robotics) là lĩnh vực lớn có sự góp mặt của machine learning. Nếu ta có một nhân vật cần di chuyển và tránh các chướng ngại vật trong game. Machine learning có thể học và giải quyết công việc này thay bạn. Một kỹ thuật phổ biến được áp dụng trong trường hợp này là Học tăng cường(Reinforcement learning). Ở đó, máy sẽ học tăng cường với mục tiêu là giải quyết nhiệm vụ trên. Học tăng cường là tiêu cực nếu nó va phải chướng ngại vật, là tích cực nếu nó chạm tới đích.
* Ưu điểm và hạn chế của Học có giám sát và không giám sát
  + Học có giám sát:
    - Ưu điểm :
      * Học có giám sát cho phép bạn thu thập dữ liệu hoặc tạo đầu ra dữ liệu từ trải nghiệm trước đó
      * Giúp bạn tối ưu hóa tiêu chí hiệu suất bằng kinh nghiệm
      * Học máy được giám sát giúp bạn giải quyết các loại vấn đề tính toán trong thế giới thực.
      * Độ chính xác cao
    - Hạn chế :
      * Ranh giới quyết định có thể được tập trung quá mức nếu tập huấn luyện của bạn không có ví dụ mà bạn muốn có trong một lớp.
      * Bạn cần chọn nhiều ví dụ hay từ mỗi lớp trong khi bạn đang đào tạo trình phân loại.
      * Phân loại dữ liệu lớn có thể là một thách thức thực sự.
      * Đào tạo cho việc học có giám sát cần rất nhiều thời gian tính toán.
  + Học không có giám sát:
    - Ưu điểm :
      * Cải thiện trải nghiệm người dùng sản phẩm và kiểm tra các hệ thống để đảm bảo chất lượng.
      * Cung cấp một con đường khám phá để xem dữ liệu, cho phép các doanh nghiệp xác định các mẫu với khối lượng lớn dữ liệu nhanh hơn khi so sánh với quan sát thủ công
    - Hạn chế
      * Tính toán phức tạp do khối lượng lớn dữ liệu đào tạo
      * Ít chính xác hơn
      * Người dùng phải hiểu và ánh xạ đầu ra thu được với các nhãn tương ứng.
      * Thời gian đào tạo dài hơn

#### **2. Trình bày phương pháp học máy được sử dụng trong bài tập lớn**

* SVM là một thuật toán giám sát, nó có thể sử dụng cho cả việc phân loại hoặc đệ quy.
* https://viblo.asia/p/gioi-thieu-ve-support-vector-machine-svm-6J3ZgPVElmB

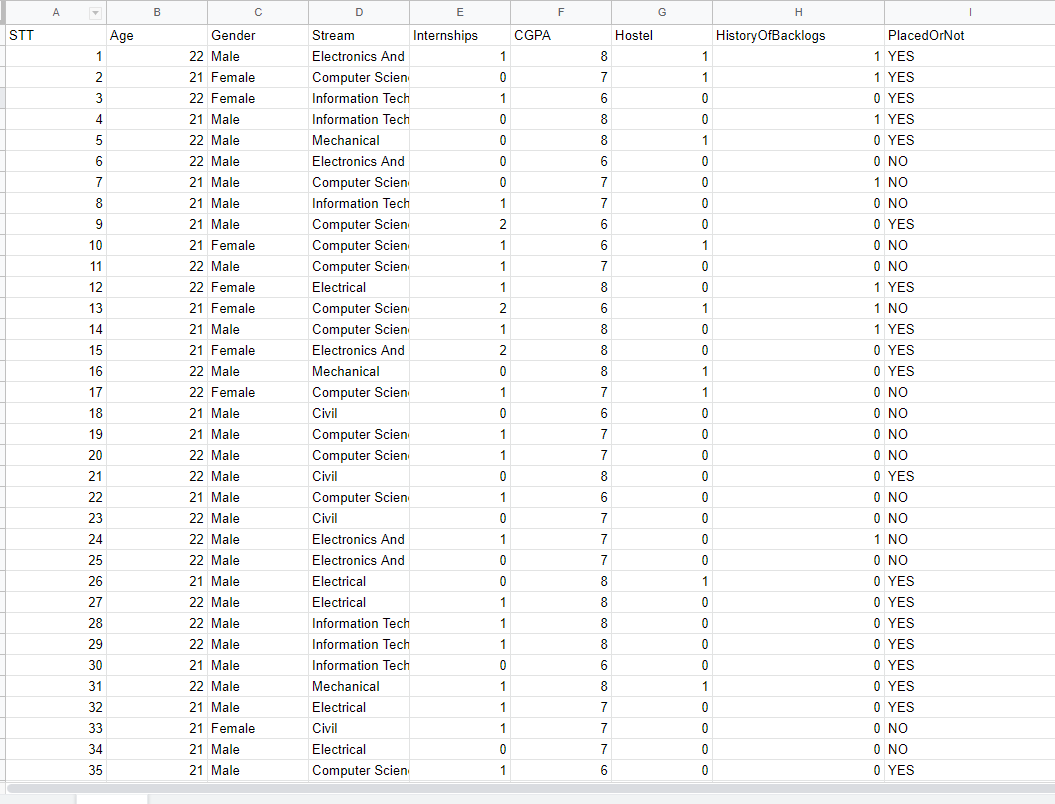
#### **3. Trình bày bài toán**

* Input: Age , Gender , Stream , Internships ,CGPA , Hostel , HistoryOfBacklogs.
* Output:PlacedOrNot .

## Phần 2: Thực nghiệm

#### **1. Mô tả tập dữ liệu**

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



* Dữ liệu được làm 2 phần bao gồm training và test
  + Training : gồm 70% dữ liệu
  + Test : gồm 30% dữ liệu

#### **2. Mô tả cách giải bài toán bằng phương pháp học máy**

**Trường hợp 2:** Nếu trong nhóm phương pháp học máy lựa chọn có phương pháp CROSS-VALIDATION

● Bước 1: Chia toàn bộ tập dữ liệu thành k=10 phần (phương pháp k-fold cross validation).

● Bước 2: Chọn lần lượt cái mẫu từ 1 đến k để làm tập test data, các phần còn lại làm data train.

○ Lần 1: Chọn tập mẫu 1 để là test data, mẫu còn lại để làm train data .

○ Lần 2: Chọn tập mẫu 2 để là test data, mẫu còn lại để làm train data .

○ ...

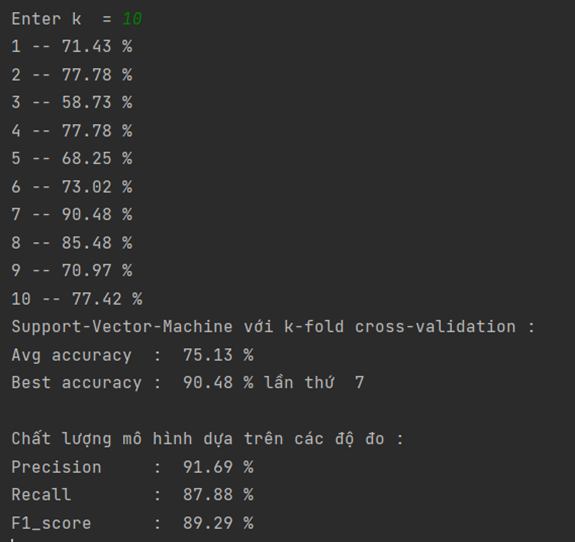
○ Lần k: Chọn tập mẫu k để là test data, mẫu từ 1 -> (K-1) là để train data.

● Bước 3: Chọn mô hình có có tỷ lệ chính xác cao nhất và có tỷ lệ sai thấp nhất .

#### **3. Đánh giá mô hình**

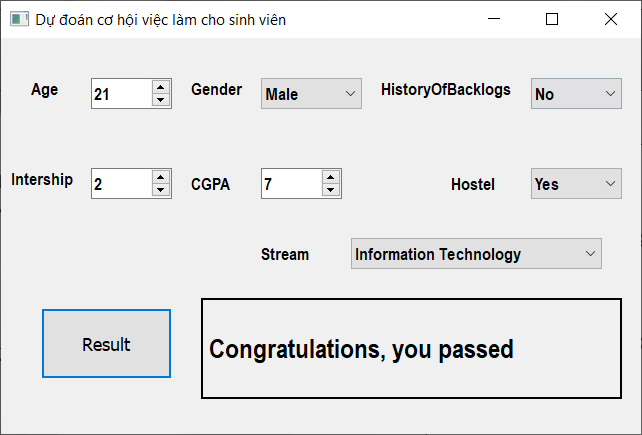
Dùng tập test data để đánh giá chất lượng của mô hình được lựa chọn theo các độ đo:

- Sau khi sử dụng phương pháp SVM ta chọn ra tập training data và test data có độ chính xác cao nhất là 90.48 %



#### **4. Mô tả các chức năng của chương trình**

* Nhóm đã thành công chạy thuật toán .
* Nhóm đã tạo giao diện để người dùng có thể nhập thông tin và trả ra kết quả dự đoán :
  + Age
  + Gender
  + Internship
  + HistoryOfBacklog:
  + CGPA
  + Hostel
  + Stream
* Result để trả ra kết quả liệu sinh viên với các thông tin như đã nhập có được nhận vào làm hay không



## Kết luận

* Xây dựng được mô hình bài toán dự đoán cơ hội việc làm cho sinh viên năm cuối từ phương pháp SVM
* Có giao diện cho phép nhập số liệu và hiển thị kết quả dự đoán.

# **Tài liệu tham khảo**

1. <https://machinelearningcoban.com/>

2. Slide bài giảng TS.Nguyễn Thị Kim Ngân.

3. <https://www.kaggle.com/tejashvi14/engineering-placements-prediction>

4.https://trituenhantao.io/kien-thuc/svm-qua-kho-hieu-hay-doc-bai-nay/?fbclid=IwAR2HLGRmdwWlkUqncWUyELZpaT5H3RZYvVgdx6t3ZSA0NS84tiACoG1MECM

#### 

#### 

#### 

### 