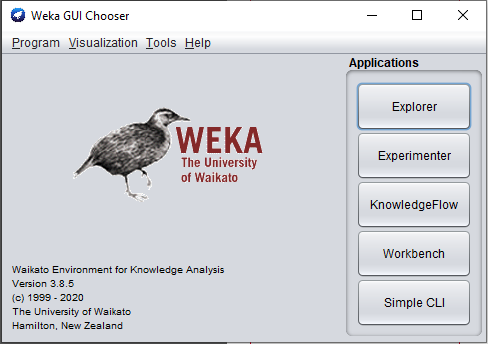
**HỌC WEKA**

**Weka là gì?**

Weka là gói phần mềm cung cấp cho người dùng các công cụ học thuật thường dùng để khai thác dữ liệu hoặc phát triển các mô hình học thuật, bao gồm các thuật toán thường được áp dụng trực tiếp vào các bộ dữ liệu hoặc đoạn code Java

 Kết quả sẽ được lưu theo các file ARFF hoặc CSV hoặc theo dạng cơ sở dữ liệu JDBC.



Explorer làm gì?

Tiền xử lý dữ liệu

Phân lớp

Phân cụm

Khai phá luật kết hợp

Lựa chọn thuộc tính

Trực quan hóa

Experimenter?

Thiết kế các thí nghiệm

Lựa chọn các thuật toán và tập dữ liệu

Chạy thí nghiệm

So sánh các kết quả

KnowledgeFlow?

Thiết kế quá trình khai phá dữ liệu một cách trực quan

Từ xử lí dữ liệu – chạy mô hình – trình bày kết quả

Workbench?

simpleCLI?

II Tập tin .ARFF là gì?

TẠI SAO CẦN CHUYỂN FILE .CSV SANG .ARFF?

định dạng chuẩn file dữ liệu của Weka la file ARFF (Attribute Relation File Format)

* Mô tả vật thể trong không gian n chiều
* Gồm 2 phần

1 header: bắt đầu từ @realation

2 phần data: bắt đầu từ @data

III Sử dụng weka trong các thuật toán

k- fold Cross -validation là gì?

 là một kỹ thuật lấy mẫu để đánh giá mô hình học máy trong trường hợp dữ liệu không được dồi dào cho lắm

Tham số quan trọng trong kỹ thuật này là k, đại diện cho số nhóm mà dữ liệu sẽ được chia ra. Vì lý do đó, nó được mang tên k-fold [cross-validation](https://trituenhantao.io/tu-dien-thuat-ngu/cross-validation/). Khi giá trị của k được lựa chọn, người ta sử dụng trực tiếp giá trị đó trong tên của phương pháp đánh giá. Ví dụ với k=10, phương pháp sẽ mang tên **10-fold**[cross-validation](https://trituenhantao.io/tu-dien-thuat-ngu/cross-validation/)**.**

Kỹ thuật này thường bao gồm các bước như sau:

1. Xáo trộn [dataset](https://trituenhantao.io/tu-dien-thuat-ngu/dataset/) một cách ngẫu nhiên
2. Chia [dataset](https://trituenhantao.io/tu-dien-thuat-ngu/dataset/) thành k nhóm
3. Với mỗi nhóm:
   1. Sử dụng nhóm hiện tại để đánh giá hiệu quả mô hình
   2. Các nhóm còn lại được sử dụng để huấn luyện mô hình
   3. Huấn luyện mô hình
   4. Đánh giá và sau đó hủy mô hình
4. Tổng hợp hiệu quả của mô hình dựa từ các số liệu đánh giá

**k=10**: Giá trị của k được gán cố định bằng 10, một giá trị thường được sử dụng và được chứng minh là cho sai số nhỏ, phương sai thấp (thông qua thực nghiệm

tóm lại : dữ liệu train sẽ được chia làm k(10) phần để đánh giá. Mỗi lần ta dung 1 phần để đánh giá còn lại k-1(9) phần làm dữ liệu train. Cuối cùng lấy tổng trung bbinhfcoongj của các lần đánh giá để đưa ra giá trị cuối cùng.