ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

BÁO CÁO LAB 3
CLASSIFICATION & CLUSTERING

Bộ môn: Khai thác dữ liệu và ứng dụng
Giảng viên hướng dẫn: Dương Nguyễn Thái Bảo

2020 - 2021

MỤC LỤC

I. THÔNG TIN NHÓM

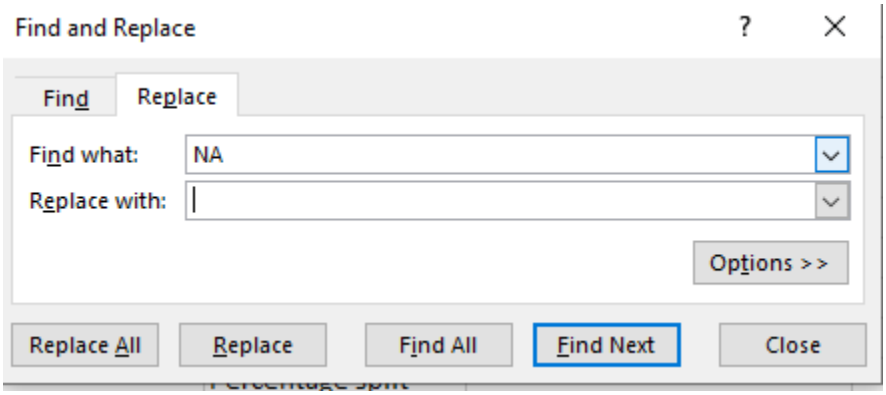
Họ tên	MSSV	Công việc
Lê Hoàng Phương Nhi	18120496	Weka Explorer, K-means
Lê Thị Như Quỳnh	18120530	Weka Experimenter, K-medoids

Mức độ hoàn thành đồ án: 100%

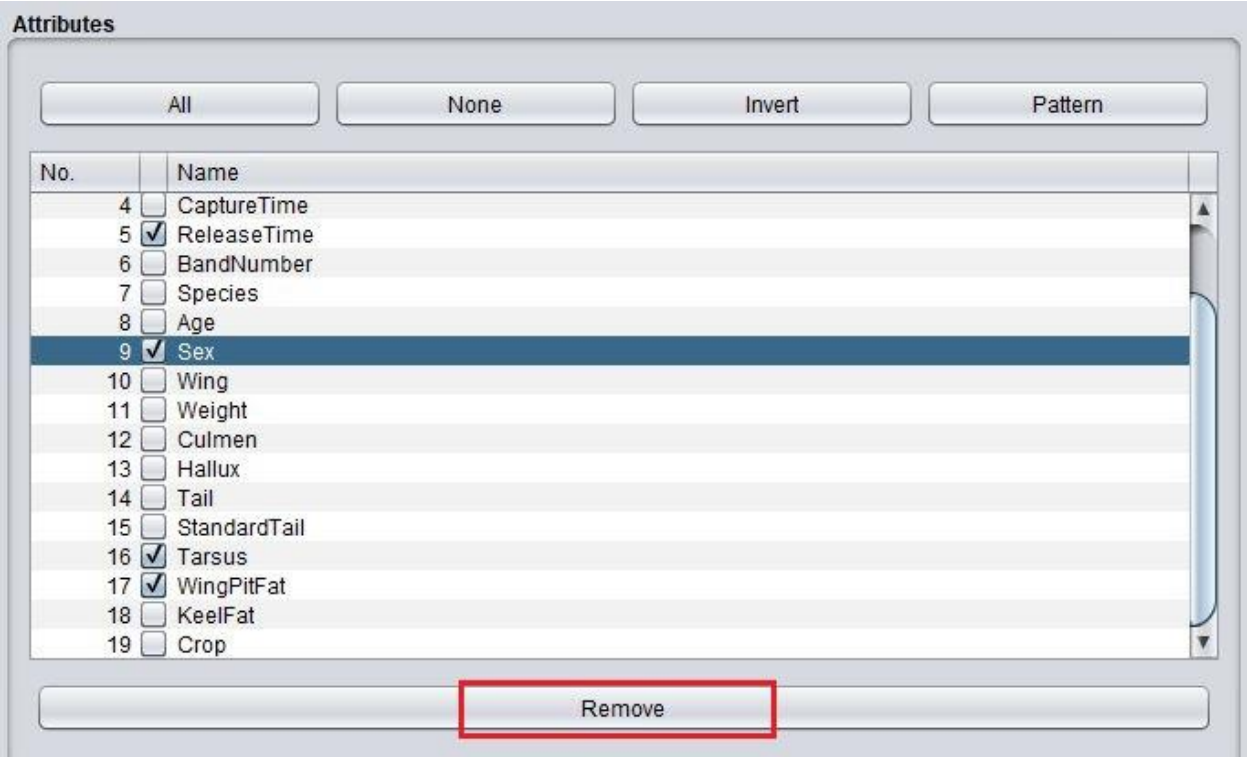
II. BÁO CÁO

1. Tiền xử lý dữ liệu

- Quan sát file hawks.csv ta thấy các giá trị NA là các giá trị thiếu, vì vậy trước khi đưa file csv này vào Weka để tiền xử lý và thực hiện yêu cầu đề bài thì ta sẽ xóa các giá trị NA này trước bằng cách sử dụng Find and Replace

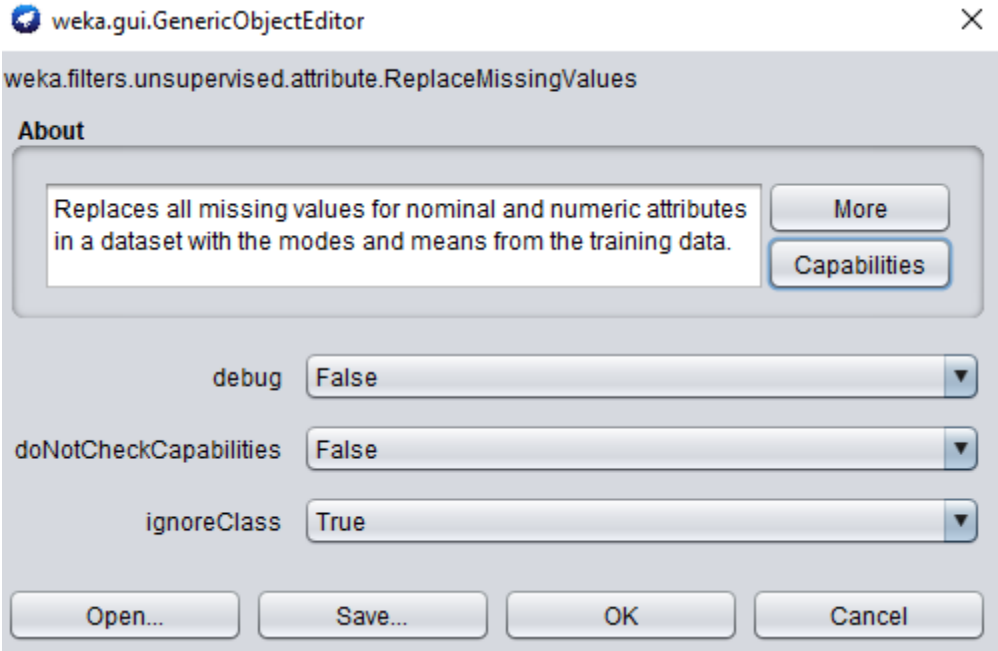


- Tiến hành đọc dữ liệu vào Weka. Quan sát trên Weka ta thấy các thuộc tính: ReleaseTime, Sex, Tarsus, WingPitFat thiếu quá nhiều giá trị (>60%) và cũng nhận thấy 4 thuộc tính này không ảnh hưởng đến quá trình phân lớp vì vậy ta có thể loại bỏ các thuộc tính này. Bằng cách: Ở khung Attributes, ta check các thuộc tính này sau đó chọn Remove



- Tập dữ liệu chứa những giá trị thiếu (missing values) vì vậy ta cần xử lý các giá trị này. Sử dụng bộ lọc ReplaceMissingValues của Weka để thực hiện điều này.

Ở khung Filter -> Choose -> filters -> unsupervised -> attribute -> ReplaceMissingValues. Và chỉnh các tham số như sau:



Sau đó nhấn Apply

2. Trả lời câu hỏi và quan sát của nhóm

- Phương pháp phân lớp nào thường cho kết quả cao nhất?

Phương pháp phân lớp NaïveBayesSimple thường cho kết quả cao nhất

- Phương pháp nào không thực hiện tốt và tại sao?

Phương pháp ID3 không thực hiện tốt. Vì:

+ Phương pháp này sẽ gặp phải vấn đề về dữ liệu quá khớp (overfitting). Đây là hiện tượng mô hình ghi nhớ quá tốt dữ liệu huấn luyện và phụ thuộc vào nó, việc này khiến cho mô hình không thể tổng quát hóa các quy luật để hoạt động với dữ liệu chưa từng được chứng kiến.

+ Chỉ hỗ trợ cho dữ liệu dạng tính, không hỗ trợ với dữ liệu liên tục.

+ Sử dụng độ đo Information Gain và phương pháp này lại ưu tiên những thuộc tính có số lượng lớn các giá trị mà ít xét tới những giá trị nhỏ hơn.

+ Không đảm bảo xây dựng được cây tối ưu

- Tại sao ta sử dụng phiên bản đã rời rạc hóa của tập dữ liệu nếu tập dữ liệu đã được rời rạc hóa?

Vì việc rời rạc hóa tập dữ liệu sẽ giúp làm giảm số lượng giá trị cho các thuộc tính liên tục điều này sẽ có thể làm tăng cường hiệu năng. Và việc rời rạc hóa dữ liệu còn đóng vai rò quan trọng trong việc biểu diễn tri thức vì chúng dễ dàng được xử lý cũng như thể hiện tri thức trực quan hơn.

- Việc rời rạc hóa và cách rời rạc hóa có ảnh hưởng đến kết quả phân lớp hay không, nếu có thì ảnh hưởng thế nào?

Việc rời rạc hóa và cách rời rạc hóa có ảnh hưởng đến kết quả phân lớp. Nó sẽ giúp tăng hiệu suất, hiệu năng của quá trình phân lớp và sẽ mang đến kết quả khả quan hơn.

- Chiến lược nào trong ba chiến lược đánh giá đã đánh giá quá cao (overestimate) độ chính xác và tại sao?

Chiến lược Use training set đánh giá quá cao độ chính xác. Vì chiến lược này sẽ đánh giá độ chính xác trên tập Training set mà không sử dụng tập Test set. Mà tập Test được sử dụng để đánh giá độ hiệu quả của mô hình, mức độ chính xác trong việc phân loại dữ liệu. Vì vậy độ chính xác của chiến lược này được đánh giá quá cao và không đáng tin cậy.

- Chiến lược nào đánh giá thấp (underestimate) độ chính xác và tại sao?

Chiến lược Percentage split đánh giá thấp độ chính xác. Vì chiến lược này sẽ hiệu quả hơn với tập dữ liệu đủ lớn. Với tập dữ liệu nhỏ, việc phân chia tập dữ liệu thành tập Training và tập Test theo tỉ lệ không phù hợp sẽ ảnh hưởng đến độ chính xác.

III. TÀI LIỆU THAM KHẢO

- https://nguyenvanhieu.vn/thuat-toan-phan-cum-k-means/
- https://pythonprogramming.net/k-means-from-scratch-2-machine-learning-tutorial/
- https://machinelearningmastery.com/design-and-run-your-first-experiment-in-weka/
- http://madhugnadig.com/articles/machine-learning/2017/03/04/implementing-k-means-clustering-from-scratch-in-python.html

- https://bigdatauni.com/vi/tin-tuc/phuong-phap-danh-gia-mo-hinh-phan-loai-classification-model-evalutation.html
- https://itguru.vn/blog/10-cau-hoi-phong-van-thuong-gap-trong-nganh-machine-learning-hoc-may-phan-1/
- https://trituenhantao.io/tu-dien-thuat-ngu/overfitting/
- https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/