Question 1:

**Trình bày về Random Tree? Tại sao phần thực nghiệm có RandomTree và NaiveBayes mà chương 2 không trình bày chỉ trình bày SVM.**

Trả lời:

1. RandomTree

Cây ngẫu nhiên là một tập hợp các cây quyết định cá nhân, mỗi cây được tạo ra từ các mẫu khác nhau và tập con của dữ liệu huấn luyện.  Ý tưởng đằng sau việc gọi các cây quyết định này là đối với mỗi giá trị được phân loại, một số quyết định được thực hiện theo thứ tự bậc quan trọng. Khi tạo đồ thị cho một giá trị, nó trông giống như một nhánh. Khi phân loại tập dữ liệu, các cành hình thành một cây. Phương pháp này được gọi là cây ngẫu nhiên bởi vì ta phân loại bộ dữ liệu một số lần dựa trên một tập hợp các giá trị huấn luyện ngẫu nhiên, do đó dẫn đến nhiều cây quyết định.

1. Bởi vì SVM là một phương pháp phân loại dựa trên nguyên tắc tối thiểu rủi ro cấu trúc, sao cho lỗi trên tập kiểm tra là nhỏ nhất và các công trình nghiên cứu đi trước cho thấy rằng SVM đưa ra độ chính xác phân lớp tốt nhấp so với phương pháp khác. Ngoài ra SVM là phương pháp sử dụng không gian vector đặc trưng lớn nhất ( hơn 10000 chiều) trong khi các phương pháp khác có số chiều thấp hơn.

Do hạn chế về mặt thời gian và để đánh giá thêm hiệu suất của mô hình dự đoán nên ở phần thực nghiệm tác giả tiến hành luyện tập dữ liệu thêm trên 2 mô hình học máy phổ biến khác là RDT và NVB. Tác giả sử dụng Công cụ Weka để tiến hành đánh giá.

**Question 2:**

**SVM có thể so sánh với các phương pháp dự đoán giới tính khác được không ? Tại sao ?**

Trả lời:

SVM ngoài so sánh với 2 thuât toán RDT và NVB thì có thể so sánh với nhiều thuật toán khác, tuy nhiên

như đã biết, phân lớp dữ liệu là một tiến trình đưa các dữ liệu chưa biết nhãn vào các lớp dữ liệu đã biết nhãn tương ứng. Mỗi nhãn được xác định bởi một số tập dữ liệu mẫu của nhãn đó.

* Chúng ta có thể thấy các thuật toán phân lớp hai lớp đều có đặc điểm chung là yêu cầu dữ liệu phải được biểu diễn dưới dạng vector đặc trưng, tuy nhiên các thuật toán khác đều phải sử dụng các ước lượng tham số và ngưỡng tối ưu trong khi đó thuật toán SVM có thể tự tìm ra các tham số tối ưu này.
* Trong các phương pháp thì SVM là phương pháp sử dụng không gian vector đặc trưng lớn nhất (hơn 10.000 chiều) trong khi đó các phương pháp khác có số chiều bé hơn nhiều
* Trong công trình của mình năm 1999, Joachims [20] đã so sánh SVM với các thuật toán học máy khác. Các kết quả cho thấy rằng SVM đưa ra độ chính xác phân lớp tốt nhất khi so sánh với các phương pháp khác.
* Luận văn đi sâu và áp dụng SVM vào bài toán dự đoán giới tính người dung Internet.

**Question 3: Tại sao độ chính xác về Nam thấp hơn nữ**

Trả lời:

Tập dữ liệu PAKDD 2015 sau khi được tác giả phân tích và đưa ra số liệu thống kê thu gọn ở hình 2.10 trong luận văn cho thấy:

1. Có sự mất cân bằng trong dữ liệu, tỉ lệ các bản ghi có nhãn là nữ nhiều hơn so với các bản ghi có nhãn là nam trong tập dữ liệu huấn luyện.
2. Các đặc trưng hiện sự quan tâm, sở thích dựa trên các yếu tố về giới tính để dự đoán giới tính là Nam có độ nhiễu cao, không phân loại rõ rệt như giới tính Nữ.