**Đề xuất đề tài bài tập lớn môn trí tuệ nhân tạo.**

**Nhóm thực hiện:**

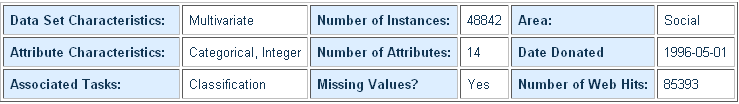
* Lê Minh Nghĩa- Lớp CNPM K51.
* Phạm Tuấn Long- Lớp CNPM K51.
* Bùi Trung Hiếu- Lớp CNPM K51.

1. **Bài toán.**
2. **Mô tả và đặc điểm.**

Bài toán phân loại dân cư từ một bộ dữ liệu có sẵn nhằm xác định nhóm người có thu nhập lớn hơn hoặc bằng 50000$ và nhóm có thu nhập dưới 50000$.

Bài toán sử dụng bộ dữ liệu Adult, được cung cấp miễn phí tại địa chỉ: <http://archive.ics.uci.edu/ml/datasets/Adult>

Đặc điểm của bộ dữ liệu :



Bài toán có ý nghĩa thực tế cao. Những kiến thức và kỹ thuật được áp dụng để giải quyết bài toán có thể áp dụng cho sự phân loại dân cư, khách hàng trong nhiều ngành và lĩnh vực khác nhau.

1. **Mô tả Input & Output**

**Bài toán 1:**

**Input:** Tập dữ liệu mẫu (training data).

**Output:** Mô hình để sử dụng phân loại dân cư.

**Bài toán 2:**

**Input:** Một chuỗi bản ghi dữ liệu về những người nào đó.

**Output:** Đưa ra dự đoán là từng người trong đó thuộc nhóm nào (thu nhập trên $50.000/năm) và tính phần trăm chính xác.

1. **Phương pháp giải quyết.**

Bài toán được giải quyết dựa trên thuật toán NBTree (Hybrid Decision Tree) của tác giả: Ron Kohavi: <http://www.aaai.org/Papers/KDD/1996/KDD96-033.pdf>

Ý tưởng của tác giả là kết hợp giữa việc phân loại theo cây quyết định sử dụng các thuật toán ID3 hoặc CD5, và phương pháp phân loại theo xác suất của phương pháp Bayes khi xác định một node lá thuộc về nhóm nào đó.

Bằng kết quả thực nghiệm tác giả đã chỉ ra độ chính xác cao của thuật toán.

**Mục tiêu :**

* **Mức một:** áp dụng thành công thuật toán NBTree với bài toán phân loại bộ dữ liệu adult của tác giả.
* **Mức hai:** Cài đặt các thuật toán : CD5, Naïve Bayes để so sánh với thuật toán NB Tree.

1. **Kết quà kỳ vọng.**

* Cài đặt thành công các thuật toán.
* Đánh giá độ phức tạp của thuật toán.
* Đánh giá so sánh tính chính xác giữa các thuật toán.
* Đánh giá độ lớn của bộ dữ liệu training cần thiết.