

BÀI TOÁN COMPUTATIONAL THINKING

Đề bài. Dự đoán khả năng chữa khỏi/tử vong của bệnh nhân covid-19

1. Abstraction.

Input: Dữ liệu 10000 bệnh nhân về tuổi tác, giới tính, bệnh án, tiền sử bệnh án, tình hình dịch bệnh tại khu vực, ...

Output: Khả năng chữa khỏi/ tử vong

2. Pattern Recognition.

Bài toán xử lý dữ liệu và trực quan hóa dữ liệu

Binary classification - Phân lớp nhị phân $\{0;1\} \Rightarrow$ Xác suất target1 = 0.3

3. Decomposition.

- Tiền xử lý dữ liệu.
 - + Phân tích dữ liệu
 - + Lọc dữ liệu nhiễu
 - + Xử lý dữ liệu bị thiếu
 - Mô hình trực quan dữ liệu: Vẽ mô hình 2D, 3D các dữ liệu
 - Tìm mối liên hệ của các dữ liệu: Tính các xác suất sau.
 - + Xác suất bệnh nhân khỏi bệnh trong từng cột mốc độ tuổi, giới tính
 - + Mắc bệnh trước đó
 - + Khu vực nhiều bệnh nhân, ít bệnh nhân
 - + Xác suất có điều kiện
- \Rightarrow Xem dữ liệu nào chiếm trọng số lớn hơn

4. Algorithm Design.

- Xử lí dữ liệu:

- + Xem thông tin dữ liệu (kích thước, nội dung,...)
- + Tách các dữ liệu thành các thuộc các loại như sau: chữ, số, phân loại
- + Chọn các dữ liệu nó khác biệt nhất trong tập dữ liệu (nó quá lớn hoặc quá nhỏ) rồi loại bỏ nó
- + Ở dữ liệu bị thiếu và điền các trung bình của các feature rồi điền vào dữ liệu bị thiếu

- Mô hình trực quan dữ liệu:

- + Vẽ mô hình phân phối mean, max, min, std, của mỗi hàng mỗi cột dữ liệu
 - + Vẽ biểu đồ mật độ của target 0 và target 1
 - + Vẽ biểu đồ các feature có correlation cao nhất với target
- Sử dụng Logistic Regression để thực hiện tính toán xác suất, huấn luyện mô hình

5. Testing/ Evaluation: F1-Score