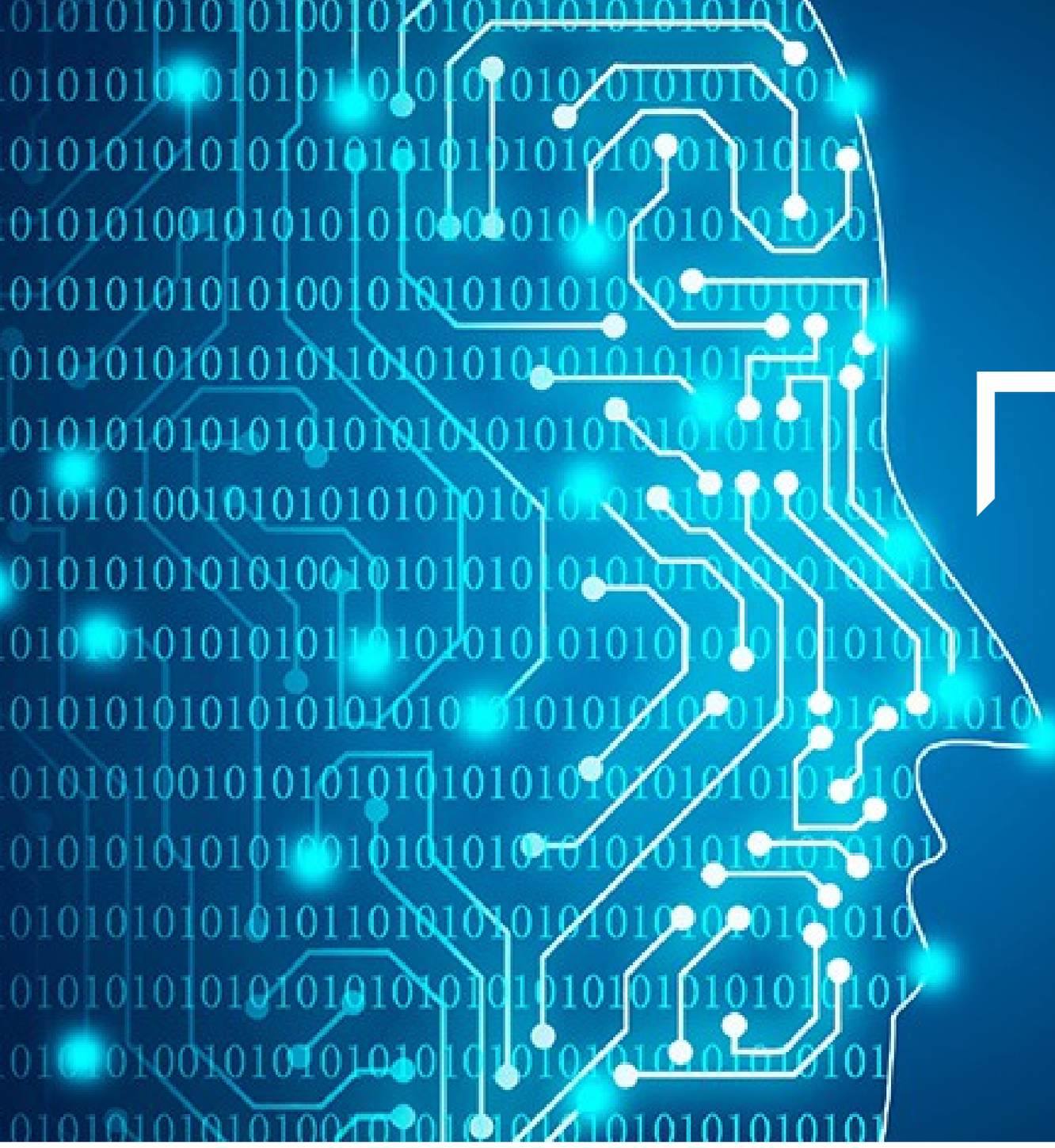


CẤU TRÚC DỮ LIÊU & GIẢI THUẬT

NHÓM 5

GIỚI THIỆU

1 Dữ liệu và khoa học dữ liệu

2 Decision Tree

3 Random Forest

4 Ứng dụng thực tế xây dựng mô hình machine learning đơn giản để dự đoán giá nhà

5 Mô phỏng

01

Dữ liệu và khoa học dữ liệu

Dữ liệu là tập hợp thông tin bao gồm các số, từ hoặc hình ảnh, được chia làm dữ liệu thô và dữ liệu đã được xử lý.

Dữ liệu thô

là các số, ký tự, hình ảnh, ký hiệu, đại lượng vật lý và thường được tiếp tục xử lý bởi con người hoặc đưa vào máy tính.

Dữ liệu thô mang tính tương đối

Dữ liệu đã được xử lí ở bước này có thể được gọi là dữ liệu thô ở bước tiếp theo.

Dữ liệu đã được xử lí

Là những dữ liệu được thu thập để chuyển đổi sang dạng mong muốn. Và phải được xử lí theo từng bước như lưu trữ, sắp xếp, xử lý, phân tích, trình bày.

Big Data

- là tập hợp dữ liệu có khối lượng lớn và phức tạp mà các phần mềm xử lý dữ liệu truyền thống không thể thu thập, quản lý và xử lý trong một khoảng thời gian ngắn.
- Bao gồm dữ liệu có cấu trúc, không có cấu trúc và bán cấu trúc

Đặc trưng của Big Data

1. Volume: Khối lượng dữ liệu lớn
2. Variety: Đa dạng các loại dữ liệu
3. Velocity: Tốc độ xử lý và phân tích dữ liệu

Khoa học dữ liệu

là khoa học về việc quản trị và phân tích dữ liệu, trích xuất các giá trị từ dữ liệu để tìm ra các hiểu biết, các tri thức hành động, các quyết định dẫn dắt hành động.

Khoa học dữ liệu bao gồm 3 phần chính

Tạo ra và quản trị dữ liệu

Phân tích dữ liệu

Chuyển kết quả phân tích thành giá trị hành động

02

Decision Tree

là cây nhị phân chia tách một cách đệ quy tập dữ liệu cho đến khi chúng ta

chỉ còn các nút lá thuần túy

Classification Decision

Cây quyết định là một thuật toán tham lam, nó chọn con đường tốt nhất, tối đa hóa thông tin thu được, nó sẽ không quay lại và thay đổi phân tách trước đó

Vì vậy tất cả các phần tách sau sẽ phụ thuộc vào phần hiện tại và điều này không đảm bảo chúng ta có được bộ phân tách tối ưu nhất nhưng có sự tham lam tìm kiếm làm cho machine learning nhanh hơn nhiều.

ENTROPY

Là thước đo lượng thông tin chứa trong một trạng thái. Tìm mức tăng thông tin tương ứng với một phép tách, chúng ta cần trừ entropy kết hợp của các nút con khỏi entropy của nút cha

- $IG = E(parent) - \sum_i \omega_i E(child_i)$

IG = E(parent) -

GINI

Dùng GINI để tính toán mức tăng thông tin, chúng ta cần kiểm tra xem mức tăng thông tin hiện tại này có lớn hơn mức tăng thông tin tối đa hay không

- $\text{Gini Index} = 1 - \sum p_i^2$ với p_i = probability of class i

Gini Index = 1 - với p_i = probability of class i

Regressor Decision

Trong hồi quy, ta sử dụng phương sai làm thước đo tạp chất giống như đã sử dụng chỉ số entropy hoặc gini trong bài toán phân loại.

- $\text{Var} = \frac{1}{n}\sum(x_i - \bar{x})^2$

  Var

Phương sai cao hơn có nghĩa là tạp chất cao hơn.

- $\text{Var Red} = \text{Var(parent)} - \sum \omega_i Var(child_i)$

  Var Red = Var(parent) -

  (Varian reduction: độ giảm)

  (Varian reduction: độ giảm)

Trọng số chỉ là kích thước tương đối của nút con đối với nút cha. Dùng Var Red với mục đích tương tự IG.

03

Random Forest

Là một tập hợp của nhiều cây quyết định ngẫu nhiên và nó ít nhạy cảm hơn

nhiều với dữ liệu

- Đào tạo một cây quyết định trên từng bộ dữ liệu được khởi động 1 cách độc lập. Chọn ngẫu nhiên một tập hợp con các đặc điểm cho từng cây và chỉ sử dụng chúng để training.

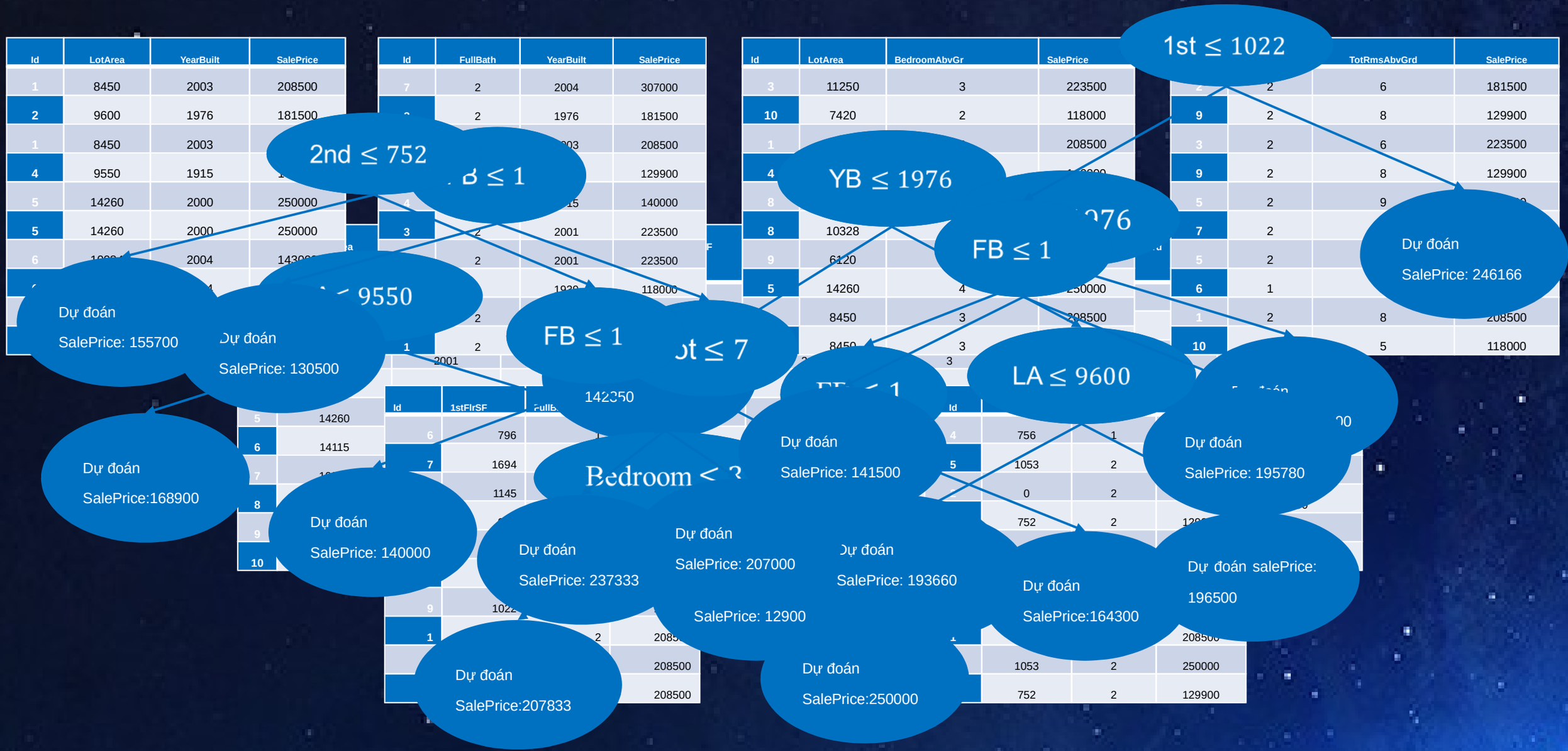
- Xây dựng cây quyết định ngẫu nhiên

- Bước đầu tiên là xây dựng bộ dữ liệu mới từ dữ liệu ban đầu. Quá trình tạo mới dữ liệu được gọi là Bootstrapping

- Kết hợp các kết quả từ nhiều mô hình (Quá trình tổng hợp)

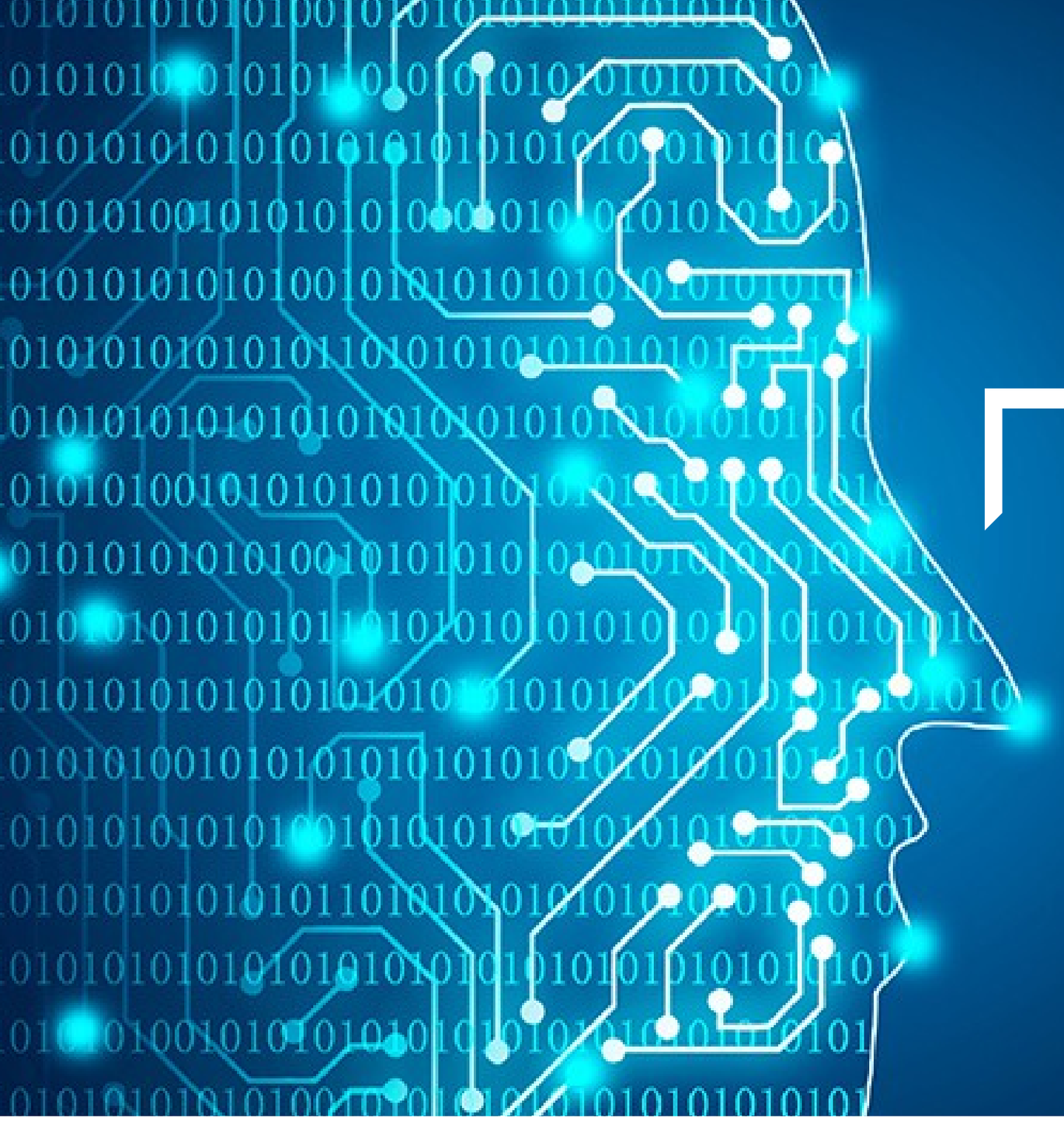
04

Ứng dụng thực tế xây dựng mô hình machine learning đơn giản để dự đoán giá nhà


1st ≤ 1022
2nd ≤ 752
YB ≤ 1976
FB ≤ 1
LA ≤ 9600
Bedroom < 3
Dự đoán
SalePrice: 155700
Dự đoán
SalePrice: 130500
Dự đoán
SalePrice: 246166
Dự đoán
SalePrice:168900
Dự đoán
SalePrice: 140000
Dự đoán
SalePrice: 141500
Dự đoán
SalePrice: 195780
Dự đoán
SalePrice: 237333
Dự đoán
SalePrice: 207000
Dự đoán
SalePrice: 193660
Dự đoán
SalePrice: 12900
Dự đoán
SalePrice:164300
Dự đoán salePrice: 196500
Dự đoán
SalePrice:207833
Dự đoán
SalePrice:250000

05

Mô phỏng



THANK YOU

Cảm ơn thầy đã lắng nghe bài thuyết trình của chúng em