**TRƯỜNG ĐẠI HỌC THỦY LỢI**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**BÀI TẬP LỚN**

**MÔN HỌC MÁY**

**ĐỀ TÀI:**

**ỨNG DỤNG THUẬT TOÁN ID3 VÀ THUẬT TOÁN PERCEPTRON LEARNING ALGORITHM(PLA) VÀO DỰ ĐOÁN GIÁ NHÀ Ở VIỆT NAM**

**Nhóm sinh viên thực hiện:**

1.Hoàng Thị Mai - 60TH1

2.Nguyễn Thị Hồng Hải - 60TH1

**Giảng viên hướng dẫn:**

PGS.TS Nguyễn Thanh Tùng

HÀ NỘI, NĂM 2021

1. **THỰC NGHIỆM**

# **Mô tả bài toán**:

# Tên bài toán : Ứng dụng thuật toán ID3 và thuật toán Perceptron learning algorithm vào dự đoán giá nhà ở Việt Nam.

# Tóm tắt công việc thực hiện của bài toán:

# + Bước 1: Thu thập dữ liệu bài toán:Dataset: <https://www.kaggle.com> .

# + Bước 2: Mô tả tập dữ liệu của bài toán:

Input (11 trá trị ): gồm

* BEDROOMS ( PHÒNG NGỦ )
* BATHROOMS ( PHÒNG TẮM )
* STORIES (LỊCH SỬ LÂU ĐỜI)
* MAINROAD ( ĐƯỜNG CHÍNH )
* GUESTROOM ( PHÒNG KHÁCH )
* BASEMENT (TẦNG HẦM)
* HOTWATERHEATING (BÌNH NÓNG LẠNH)
* AIRCONDITIONING (ĐIỀU HÒA NHIỆT ĐỘ)
* PARKING (BÃI ĐẬU XE)
* PREFAREA
* FURNITURETATUS (NỘI THẤT)

Output: dự đoán giá nhà

* PRICE\_SEGMENT (TẦM GIÁ)

# + Bước 3: Mô tả ma trận dữ liệu (X), nhãn lớp (Y)

# + Bước 4: Chia tập dữ liệu thành 2 phần: 70% dùng để huấn luyện mô hình, 30% dùng để kiểm tra sự phù hợp của mô hình.

# + Bước 5: Dùng tập dữ liệu kiểm tra để so sánh tỷ lệ mẫu được dự đoán đúng của Perceptron và ID3 trên python.

# **2. Mô tả tập dữ liệu của bài toán**:

* Dữ liệu bài toán gồm: ma trận dữ liệu (X), nhãn lớp (Y)
* Ma trận x là vector đầu vào tập hợp các mẫu dữ liệu, mỗi mẫu dữ liệu là một vetor mô tả giá nhà nhỏ hơn, lớn hơn hoặc bằng 7 tỷ là nhãn lớp y (giá trị thực của đầu ra) gồm có 11 thông tin dữ liệu:
* BEDROOMS ( PHÒNG NGỦ )
* BATHROOMS ( PHÒNG TẮM )
* STORIES (LỊCH SỬ LÂU ĐỜI)
* MAINROAD ( ĐƯỜNG CHÍNH )
* GUESTROOM ( PHÒNG KHÁCH )
* BASEMENT (TẦNG HẦM)
* HOTWATERHEATING (BÌNH NÓNG LẠNH)
* AIRCONDITIONING (ĐIỀU HÒA NHIỆT ĐỘ)
* PARKING (BÃI ĐẬU XE)
* PREFAREA
* FURNITURETATUS (NỘI THẤT)
* Bài toán có 150 mẫu dữ liệu.
* Dữ liệu bài toán:

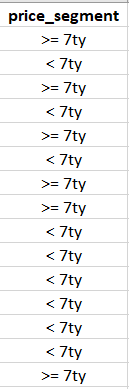
Dataset: <https://www.kaggle.com> .

# **Mô tả ma trận dữ liệu (X), nhãn lớp (Y)**

* Ma trận dữ liệu X: demo 15 mẫu dữ liệu ma trận X



* Ma trận nhãn lớp Y: demo 15 mẫu dữ liệu nhãn lớp Y



# **4. Chia tập dữ liệu thành 2 phần:**

* Tập dữ liệu có 150 mẫu dữ liệu chia: 105 mẫu dữ liệu dùng để huấn luyện mô hình, 45 mẫu dữ liệu dùng để kiểm tra sự phù hợp của mô hình.
* Dữ liệu tập huấn luyện (Training).
* Ma trận xtrain là vector đầu vào tập hợp các điểm dữ liệu, mỗi mẫu dữ liệu là một vetor mô tả giá nhà nhỏ hơn, lớn hơn hoặc bằng 7 tỷ là nhãn lớp ytrain (giá trị thực của đầu ra) gồm có 11 thông tin dữ liệu:BEDROOMS ( PHÒNG NGỦ ),BATHROOMS ( PHÒNG TẮM ), STORIES (LỊCH SỬ LÂU ĐỜI), MAINROAD ( ĐƯỜNG CHÍNH ), GUESTROOM ( PHÒNG KHÁCH ), BASEMENT (TẦNG HẦM), HOTWATERHEATING (BÌNH NÓNG LẠNH), AIRCONDITIONING (ĐIỀU HÒA NHIỆT ĐỘ), PARKING(BÃI ĐẬU XE), PREFAREA, FURNITURETATUS(NỘI THẤT).
* Tập dữ liệu huấn luyện gồm 105 mẫu dữ liệu.
* Dữ liệu tập kiểm tra (Testing)
* Ma trận xtest là vector đầu vào tập hợp các điểm dữ liệu, mỗi mẫu dữ liệu là một vector mô tả giá nhà nhỏ hơn, lớn hơn hoặc bằng 7 tỷ là nhãn lớp ytest (giá trị thực của đầu ra) gồm có 11 thông tin dữ liệu: BEDROOMS ( PHÒNG NGỦ ), BATHROOMS (PHÒNG TẮM), STORIES (LỊCH SỬ LÂU ĐỜI), MAINROAD ( ĐƯỜNG CHÍNH ), GUESTROOM ( PHÒNG KHÁCH ), BASEMENT (TẦNG HẦM), HOTWATERHEATING (BÌNH NÓNG LẠNH), AIRCONDITIONING (ĐIỀU HÒA NHIỆT ĐỘ), PARKING (BÃI ĐẬU XE), PREFAREA, FURNITURETATUS (NỘI THẤT).
* Tập dữ liệu kiểm tra gồm 45 mẫu dữ liệu.

# **5. Dùng thuật toán Perceptron và thuật toán ID3 để xây dựng mô hình phân lớp cho bài toán**

Demo chương trình

1. **ĐÁNH GIÁ**

1. **Mô tả kết quả dự đoán bằng thuật toán ID3**

* Demo 15 kết quả dự đoán từ mô hình huấn luyện so với giá trị thực tế

Dự đoán [ < 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ < 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ < 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ < 7ty ] Thực tế là [ >= 7ty ]

Dự đoán [ < 7ty ] Thực tế là [ < 7ty

Dự đoán [ < 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ < 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ < 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ >= 7ty ]

Dự đoán [ < 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ < 7ty ] Thực tế là [ >= 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ < 7ty ] Thực tế là [ >= 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

...

Thuật toán ID3 dự đoán đúng được 31 giá trị trên tổng số 45.

# **Mô tả kết quả dự đoán bằng thuật toán Perceptron learning algorithm (PLA)**

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ >= 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ >= 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ >= 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ >= 7ty ]

Dự đoán [ >= 7ty ] Thực tế là [ < 7ty ]

...

Thuật toán Perceptron dự đoán đúng được 20 giá trị trên tổng số 45

# **Nhận xét**

* Từ vào tập dữ liệu kiểm tra ứng dụng hai thuật toán ID3 và Perceptron, dựa vào mô tả kết quả dự đoán của hai thuật toán ở trên ta thấy rằng thuật toán ID3 tỷ lệ mẫu dự đoán đúng với giá trị thực tế nhiều hơn so với sử dụng thuật toán Perceptron.

=> Vậy nên sử dụng thuật toán ID3 phù hợp hơn với mô hình dự đoán của bài toán.

1. **TÀI LIỆU THAM KHẢO**

* Tham khảo slide “Học máy” của thầy PGS.TS Nguyễn Thanh Tùng
* Tài liệu file: <https://github.com/tiepvupsu/ebookMLCB> của thầy Vũ Hữu Tiệp
* Dataset:[https://www.kaggle.com/](https://www.kaggle.com/goyalshalini93/car-price-prediction-linear-regression-rfe/data?fbclid=IwAR0x2sXrpPkU5PBKFhvSrWrJL344T-Kwyx5Qtunm_z4TOkW3xHsPJ7Slg3s)