Họ tên: Nguyễn Minh Chiều

Lớp: 59TH1

MSV: 175A071203

**Đề bài**

**BÀI TẬP1:**

**1. Phương pháp Hồi quy tuyến tính**

**2. ID3**

**Bài 1:** Bảng dữ liệu về chiều cao và cân nặng của 15 người như dưới đây:

| **Chiều cao (cm)** | **Cân nặng (kg)** | **Chiều cao (cm)** | **Cân nặng (kg)** |
| --- | --- | --- | --- |
| 147 | 49 | 168 | 60 |
| 150 | 50 | 170 | 71 |
| 153 | 51 | 173 | 63 |
| 155 | 53 | 175 | 64 |
| 158 | 54 | 178 | 66 |
| 160 | 56 | 180 | 67 |
| 163 | 58 | 183 | 70 |
| 165 | 59 |  |  |

a, Thực hiện tính tay xây dựng đường hồi quy tuyến tính của cân nặng theo chiều cao? Dự đoán cân nặng của 1 người khi có chiều cao là 159cm?

b, Lập trình xây dựng đường hồi quy tuyến tính của cân nặng theo chiều cao? Dự đoán cân nặng của 1 người khi có chiều cao là 159cm? So sánh với kết quả tính tay?

**Bài làm**

Bài 1:

y: Cân nặng (kg)

x: Chiều cao (cm)

Theo đề bài ta có hàm: f(x) = w1x + w0

=> vector hệ số w = [w0, w1]T

Theo đề bài ta có:

* Vector cột y = [49;50;51;53;54;56;58;59;60;71;63;64;66;67;70]
* Ma trận X = 

\* Để tính được w ta có công thức XTXw = XTy ⇔ w=(XTX)-1 . Xty

XT:



=> XTX = 

(XTX)-1 = 

XTy = 

=> w = (XTX)-1 x XTy = 

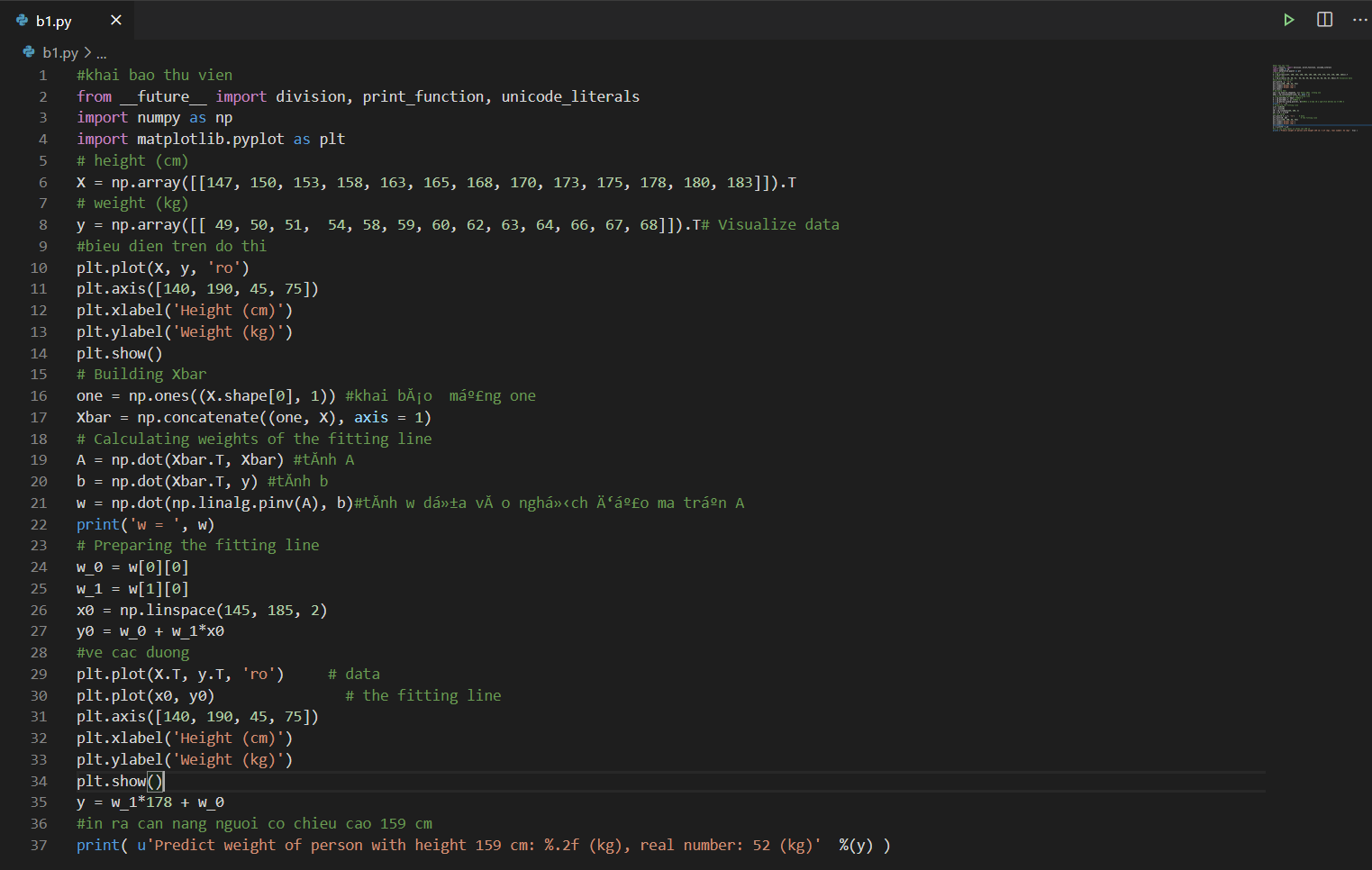
Dựa và w => w0 = -40.0862 và w1 = 0.6022

=> Ta có phương trình y = -40.0862 + 0.6022x và sai số trung bình là 0.6531

Vậy người có chiều cao là 159cm sẽ có cân nặng là:

y = -40.0862 + 0.6022 x 159 = 55.6636 55 (kg)

b



Hàm dự đoán: y ≈ f(x) = -40.0862+ 0.6022\*x

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chiều cao (cm)**  **X** | **Cân nặng (kg)**  **Y** | **Cân nặng theo f(x)**  **<code>** | **Sai số**  **| f(x)-Y |** | **Cân nặng dự đoán theo f(x)-<tính tay>** | **Sai số**  **| f(x)-Y |** |
| 147 | 49 | 48.47 | 0.53 | 48.4372 | 0.5628 |
| 150 | 50 | 50.15 | 0.15 | 50.2438 | 0.2438 |
| 153 | 51 | 51.82 | 0.82 | 52.0504 | 1.0504 |
| 155 | 53 | 52.94 | 0.06 | 53.2548 | 0.2548 |
| 158 | 54 | 54.62 | 0.62 | 55.0614 | 1.0614 |
| 160 | 56 | 55.74 | 0.26 | 56.2658 | 0.2658 |
| 163 | 58 | 57.42 | 0.58 | 58.0724 | 0.0724 |
| 165 | 59 | 58.53 | 0.07 | 59.2768 | 0.2768 |
| 168 | 60 | 60.21 | 0.21 | 61.0834 | 1.0834 |
| 170 | 71 | 61.33 | 9.67 | 62.2878 | 1.0834 |
| 173 | 63 | 63.01 | 0.01 | 64.0944 | 8.7122 |
| 175 | 64 | 64.13 | 0.13 | 65.2988 | 1.0944 |
| 178 | 66 | 65.80 | 0.20 | 67.1054 | 1.2988 |
| 180 | 67 | 66.92 | 0.08 | 68.3098 | 1.1054 |
| 183 | 70 | 68.60 | 1.40 | 70.1164 | 1.3098 |
| 159 |  | 55.18 |  | 55.6636 |  |

Sai số trung bình là 0.986

Với người có chiều cao 159cm thì ước tính cân nặng là 55.18kg

Bài 2:

a)

Gọi x: Mức lãi xuất

Gọi y: Tỉ lệ lạm phát

Như đề bài ta có hàm: f(y) = w1y + w0

Dựa vào hàm f(y) ta có vector hệ số w = [w0, w1]T

Theo đề bài ta có vector cột x = [17.5 ; 15.6 ; 9.8 ; 5.3 ; 7.9 ; 10 ; 19.2 ; 13.1]

Ta có ma trận Y = 

Để tính được w ta có công thức: YTYw = YTx ⇔ w = (YTY)-1 x YTx

Ta có YT = 

=> YTY = 

(YTY)-1 = 

YTx = 

=> w = (YTY)-1 x YTx = 

Dựa vào w => w0 = 2.7877 ,w1 = 1.0496 , sai số trung bình = 0.37067

=> Ta có phương trình

y = 2.7877 + 1.0496x

b) Ta có mức lãi xuất là 11

=> f(x) = 11

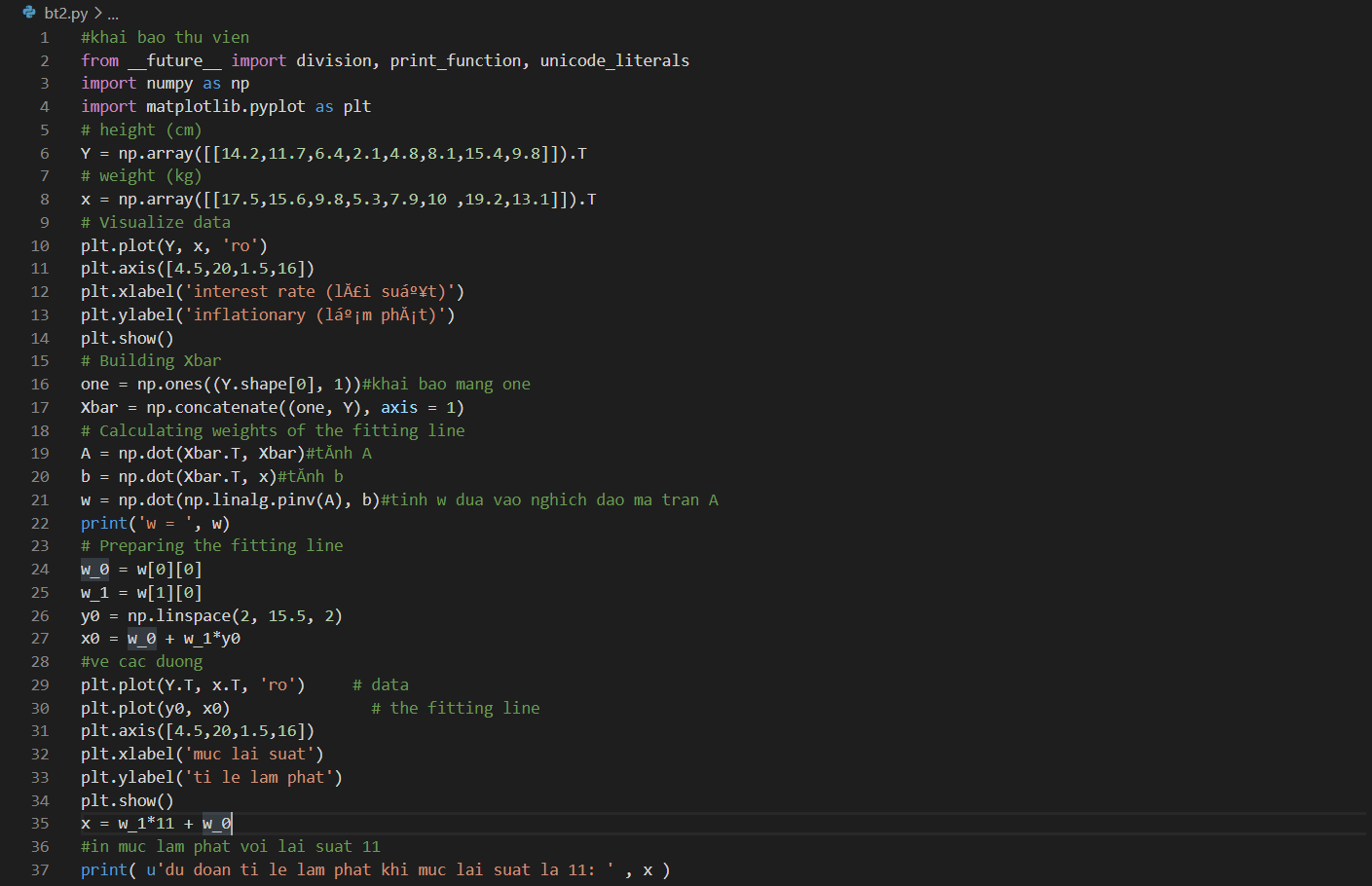
⇔ 2.7877 + 1.0496y = 11

⇔ y = (11 – 2.7877) / 1.0496

⇔ y = 7.8242

Vậy với mức lãi xuất là 11 thì tỷ lệ lạm phát là 7.8242

c.



Hàm dự đoán: x ≈ f(y) = 2.7877+ 1.0496\*y

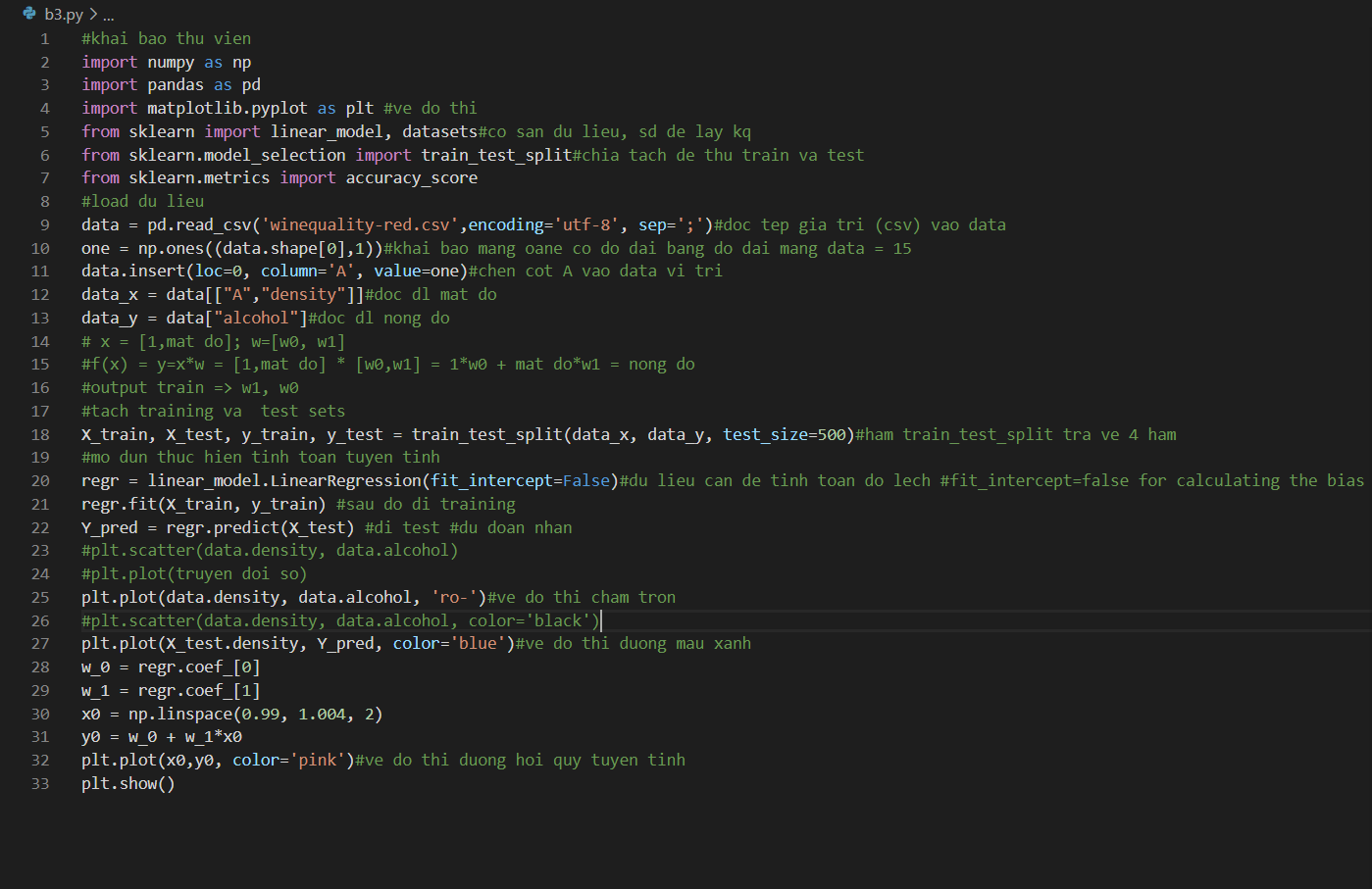
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **X**  **(lãi suất)** | **Y**  **(lạm phát)** | **Cân nặng theo f(y)**  **<code>** | **Sai số**  **| f(y)-X |** | **Cân nặng dự đoán theo f(x)-<tính tay>** | **Sai số**  **| f(x)-Y |** |
| 17.5 | 14.2 | 17.6925 | 0.1925 | 17.69202 | 0.19202 |
| 15.6 | 11.7 | 15.0684 | 0.5316 | 15.06802 | 0.53198 |
| 9.8 | 6.4 | 9.5053 | 0.2947 | 9.50514 | 0.29486 |
| 5.3 | 2.1 | 4.9919 | 0.3051 | 4.99186 | 0.30814 |
| 7.9 | 4.8 | 7.8259 | 0.0741 | 7.82578 | 0.07422 |
| 10 | 8.1 | 11.2897 | 1.2897 | 11.28946 | 1.28946 |
| 19.2 | 15.4 | 18.9520 | 0.2480 | 18.95154 | 0.24846 |
| 13.1 | 9.8 | 13.0741 | 0.0259 | 13.07378 | 0.02622 |

x = 11 => y ≈ (11 – 2.7877) / 1.0496 ≈ 7.8242

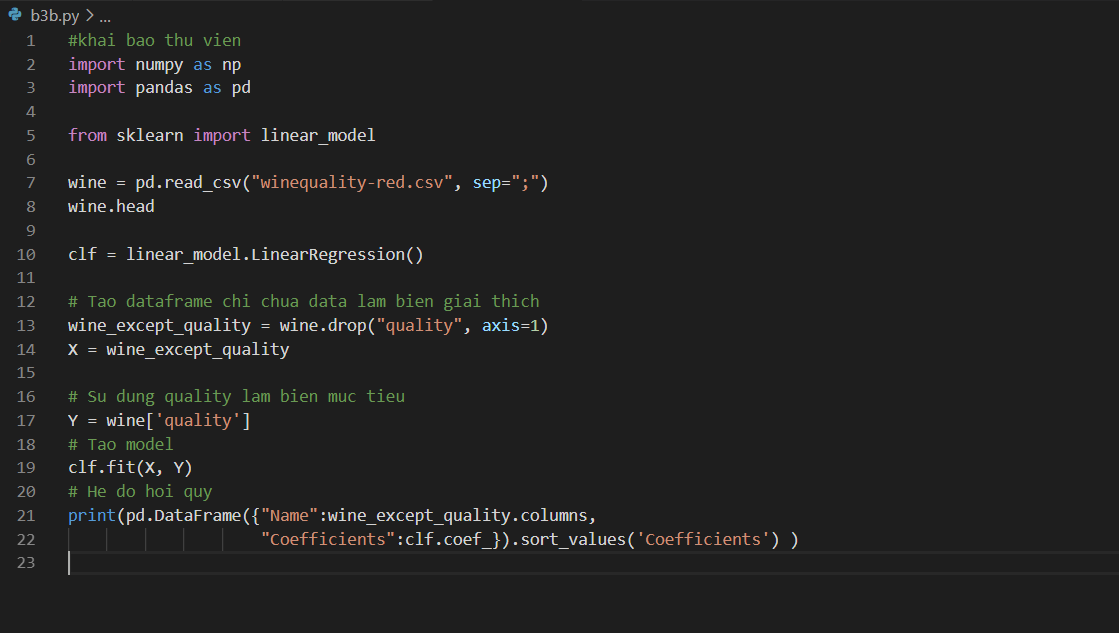
Cân nặng theo f(x)<code> y ≈ 7.8406

**Bài 3:**

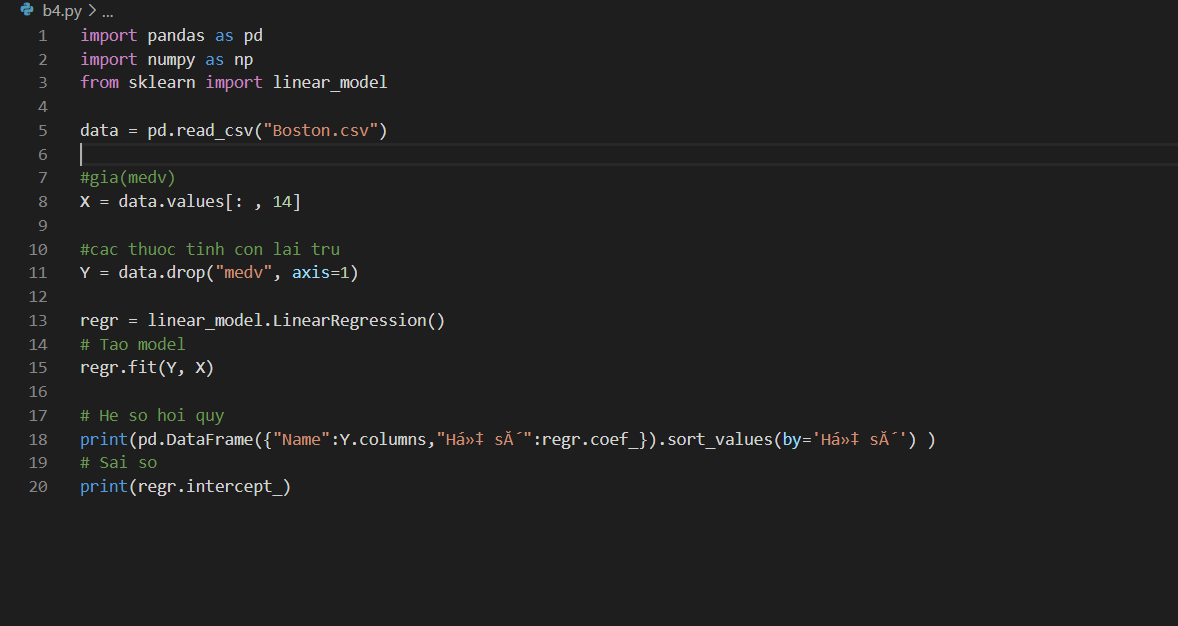
a.



b.



**Bài 4:**



**Bài 5:**

**a.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **MaHS** | **Toan** | **Ly** | **Hoa** | **ĐH** |
| 1 | Giỏi | Giỏi | Giỏi | Đỗ |
| 2 | Khá | TB | TB | Trượt |
| 3 | Giỏi | Giỏi | Khá | Đỗ |
| 4 | TB | Giỏi | Giỏi | Đỗ |
| 5 | Giỏi | Giỏi | TB | Đỗ |
| 6 | TB | Khá | Khá | Đỗ |
| 7 | Khá | Giỏi | Giỏi | Đỗ |
| 8 | TB | Khá | TB | Trượt |
| 9 | Khá | TB | Giỏi | Đỗ |
| 10 | Giỏi | TB | TB | Trượt |
| 11 | TB | TB | Giỏi | Trượt |
| 12 | Giỏi | Khá | TB | Đỗ |
| 13 | Khá | Giỏi | TB | Đỗ |
| 14 | Giỏi | TB | Khá | Đỗ |
| 15 | TB | TB | TB | Trượt |

Xét nhãn ĐH : [10+;5-]

* Thuộc tính TOAN: value TOAN = {giỏi, tb, khá}

Entropy giỏi = - = 0.65

[5+;1-]

Entropy tb = - = 0.97

[2+;3-]

Entropy khá = - = 0.81

[3+;1-]

* H(S,TOAN)= 0.65 x + 0.97 x + 0.81 x = 0.8
* Thuộc tính LY : value LY = {giỏi, tb, khá}

Entropy giỏi =

[6+;0-]

Entropy tb = - = 0.92

[2+;4-]

Entropy khá = - = 0.92

[2+;1-]

* H(S,LY)= 0.92 x + 0.92 x = 0.552
* Thuộc tính HOA : value HOA = {giỏi, tb, khá}

Entropy giỏi =

[4+;1-]

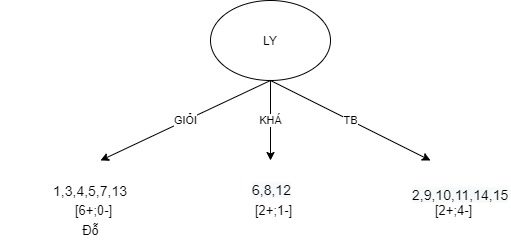
Entropy tb = - = 0.98

[3+;4-]

Entropy khá =

[3+;0-]

* H(S,LY)= 0.72 x + 0.98 x = 0.7
* Chọn nút gốc là Ly vì có H(S,Ly) min



Xét nhánh KHÁ

|  |  |  |  |
| --- | --- | --- | --- |
| MAHS | TOAN | HOA | ĐH |
| 6 | TB | Khá | Đỗ |
| 8 | TB | TB | Trượt |
| 12 | Giỏi | TB | Đỗ |

* Xét thuộc tính TOAN: value TOAN = {TB,Giỏi}

Entropy giỏi =

[1+;0-]

Entropy tb =

[1+;1-]

* H(S’’,TOAN) =
* Xét thuộc tính HOA : value HOA = {Khá,TB}

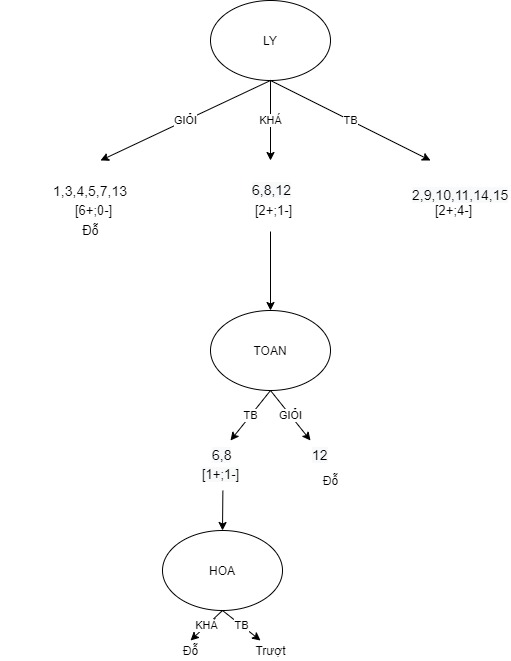
Entropy khá =

[1+;0-]

Entropy tb =

[1+;1-]

* H(S’’,TOAN) =
* Chọn thuộc tính TOAN làm gốc tiếp theo



Xét nhánh TB

|  |  |  |  |
| --- | --- | --- | --- |
| MAHS | TOAN | HOA | DH |
| 2 | Khá | TB | Trượt |
| 9 | Khá | Giỏi | Đỗ |
| 10 | Giỏi | TB | Trượt |
| 11 | TB | Giỏi | Trượt |
| 14 | Giỏi | Khá | Đỗ |
| 15 | TB | TB | Trượt |

* Xét thuộc tính TOAN : value TOAN ={Khá, giỏi,TB}

Entropy giỏi =

[1+;1-]

Entropy tb =

[0+;2-]

Entropy khá =

[1+;1-]

* H(S,LY)= + =
* Xét thuộc tính HOA : value HOA = {Khá, giỏi,TB}

Entropy giỏi =

[1+;1-]

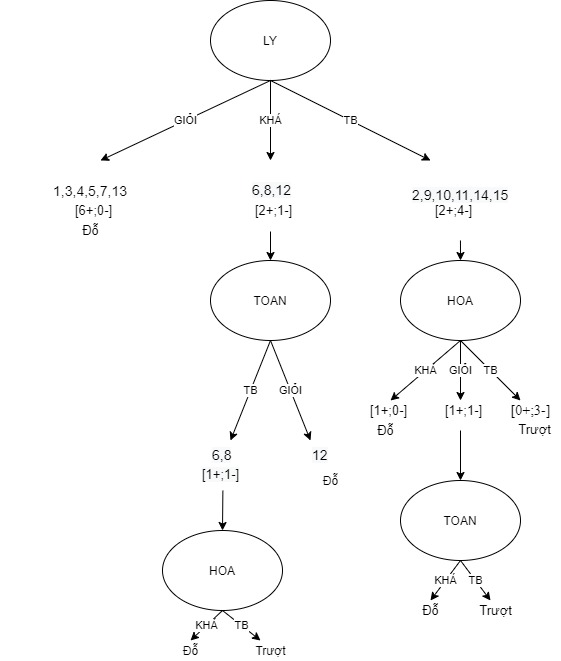
Entropy tb =

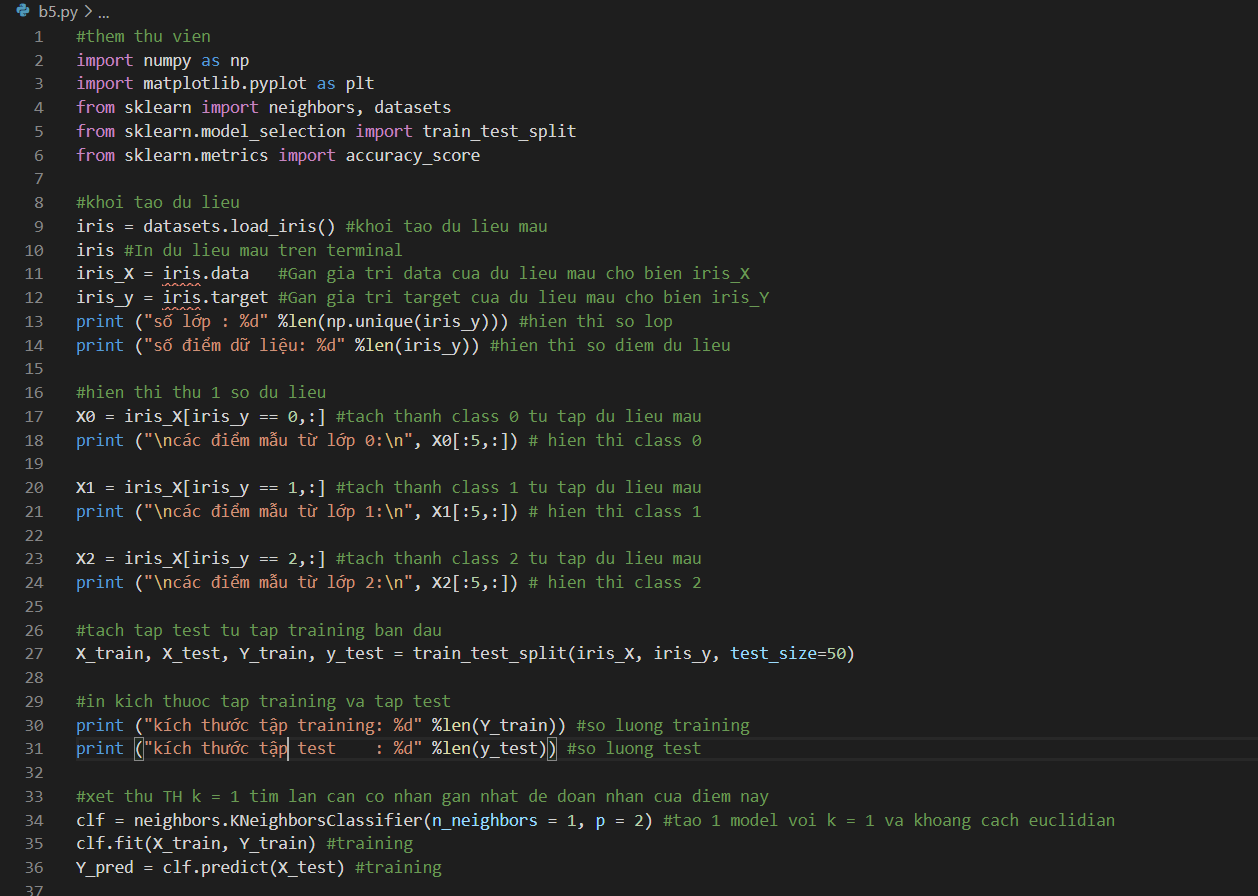
[0+;3-]

Entropy khá =

[1+;0-]

* H(S,HOA)=
* Chọn HOA làm gốc tiếp theo vì H(S’’,HOA) min



2. 

**Bài 6**. Điều tra 12 người về độ tuổi, tình trạng hôn nhân, bất động sản, thu nhập và đánh giá mức độ rủi ro tài chính được như sau:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID | Tuổi | Hôn nhân | BDS | Thu nhập | Rủi ro tài chính |
| 1 | Trẻ | Độc thân | Nhà thuê | Thấp | Không |
| 2 | Trung bình | Đã kết hôn | Nhà sở hữu | Cao | Không |
| 3 | Trẻ | Đã kết hôn | Nhà thuê | Thấp | Có |
| 4 | Trung bình | Độc thân | Nhà thuê | Thấp | Có |
| 5 | Trung bình | Đã kết hôn | Nhà sở hữu | Thấp | Có |
| 6 | Trẻ | Đã kết hôn | Nhà sở hữu | Thấp | Có |
| 7 | Trẻ | Độc thân | Nhà thuê | Trung bình | Không |
| 8 | Trung bình | Đã kết hôn | Nhà sở hữu | Cao | Không |
| 9 | Cao | Độc thân | Nhà sở hữu | Trung bình | Có |
| 10 | Cao | Đã kết hôn | Nhà sở hữu | Trung bình | Không |
| 11 | Trẻ | Độc thân | Nhà thuê | Thấp | Có |
| 12 | Trung bình | Đã kết hôn | Nhà sở hữu | Thấp | Có |

1. Yêu cầu thực hiện bằng tay:
2. Xây dựng cây quyết định bằng thuật toán ID3 sử dụng độ đo Entropy
3. Xây dựng cây CART sử dụng chỉ số Gini Index
4. Thực hiện lập trình với thư viện sklearn với phương pháp cây quyết định và rừng ngẫu nhiên để so sánh kết quả

Lời giải:

1.

a. Xây dựng cây quyết định bằng thuật toán ID3 sử dụng độ đo Entropy:

- Entropy tại root node: H(S) =

+ Xét thuộc tính Tuổi:

H(ST) =

H(STB) =

H(SC) =

H(Tuổi,S) =

+ Xét thuộc tính Hôn Nhân

H(SĐT) =

H(SĐKH) =

H(Hôn Nhân,S) =

+ Xét thuộc tính BĐS:

H(SNT) =

H(SNSH) =

H(BĐS,S) =

+ Xét thuộc tính Thu Nhập:

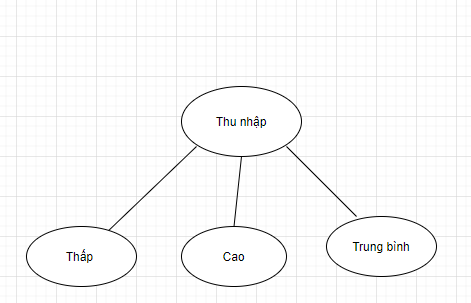
H(STNT) =

H(STNTB) =

H(STNC) =

H(Thu Nhập,S) =

* Thuộc tính cần chọn ở bước đầu tiên là Thu Nhập vì H(Thu Nhập, S) có giá trị nhỏ nhất (Information gain là lớn nhất)



- Xét Thu Nhập(Thấp)

+ Xét thuộc tính Tuổi:

H(ST) =

H(STB) =

H(Tuổi,S) =

+ Xét thuộc tính Hôn Nhân:

H(SĐT) =

H(SĐKH) =

H(Hôn Nhân,S) =

+Xét thuộc tính BĐS:

H(SNT) =

H(SNSH) =

H(BĐS,S) =

* Thuộc tính cần chọn cho phần Thu nhập(thấp) là Hôn Nhân vì H(Hôn Nhân, S) có giá trị nhỏ nhất (Information gain là lớn nhất)

-Xét Thu Nhập(Trung Bình)

+Xét thuộc tính tuổi:

H(ST) =

H(SC) =

H(Tuổi,S) =

+Xét thuộc tính BĐS:

H(SNT) =

H(SNSH) =

H(BĐS,S) =

* Có thể chọn thuộc tính Tuổi hoặc BĐS để làm nút tiếp theo của nhánh Thu Nhập(Trung Bình) vì H(Tuổi,S) = H(BĐS,S).

b. Xây dựng cây CART sử dụng chỉ số Gini Index

- Phân nhánh theo thuộc tính Tuổi:

+ Gini cho node Trẻ: 0.62 + 0.42 = 0.52

+ Gini cho node TB: 0.62 + 0.42 = 0.52

+ Gini cho node Cao: 0.52 + 0.52 = 0.5

* Trọng số Gini cho việc phân nhánh theo Tuổi:

0.52 + 0.52 + 0.5 = 0.517

-Phân nhánh theo thuộc tính Hôn Nhân:

+ Gini cho node Độc Thân: 0.62 + 0.42 = 0.52

+ Gini cho node Đã Kết Hôn: 0.572 + 0.432 = 0.5098

* Trọng số Gini cho việc phân nhánh theo Hôn Nhân:

0.52 + 0.5098 = 0.514

-Phân nhánh theo thuộc tính BĐS:

+ Gini cho node Nhà Thuê: 0.62 + 0.42 = 0.52

+ Gini cho node Nhà Sở Hữu: 0.572 + 0.432 = 0.5098

* Trọng số Gini cho việc phân nhánh theo BĐS:

0.52 + 0.5098 = 0.514

-Phân nhánh theo thuộc tính Thu Nhập:

+ Gini cho node Thấp: 0.862 + 0.142 = 0.7592

+ Gini cho node TB: 0.332 + 0.672 = 0.5578

+ Gini cho node Cao: 1

* Trọng số Gini cho việc phân nhánh theo Thu Nhập:

0.7592 + 0.5578 + 1 = 0.749

Ta thấy, Gini Score cho Thu Nhập cao hơn các thuộc tính còn lại. Do đó việc phân nhánh sẽ dựa trên thuộc tính Thu Nhập. (Chia làm 2 nhánh: Cao, Thấp và Trung Bình)

Xét nhánh Thu nhập(Thấp và TB):

-Xét thuộc tính Tuổi:

+ Gini cho node Trẻ: 0.62 + 0.42 = 0.52

+ Gini cho node TB: 1

+ Gini cho node Cao: 0.52 + 0.52 = 0.5

* Trọng số Gini cho việc phân nhánh theo Tuổi:

0.52 + 1 + 0.5 = 0.66

-Xét thuộc tính Hôn Nhân:

+ Gini cho node Độc Thân: 0.62 + 0.42 = 0.52

+ Gini cho node Đã Kết Hôn: 0.82 + 0.22 = 0.68

* Trọng số Gini cho việc phân nhánh theo Hôn Nhân:

0.52 + 0.68 = 0.6

-Xét thuộc tính BĐS:

+ Gini cho node Nhà Thuê: 0.62 + 0.42 = 0.52

+ Gini cho node Nhà Sở Hữu: 0.82 + 0.22 = 0.68

* Trọng số Gini cho việc phân nhánh theo BĐS:

0.52 + 0.68 = 0.6

Ta thấy, Gini Score cho Tuổi cao hơn các thuộc tính còn lại. Do đó việc phân nhánh Thu Nhập(thấp và trung bình) sẽ dựa trên thuộc tính Tuổi. (Chia làm 2 nhánh: Trung Bình, Trẻ và Cao)

Xét nhánh Tuổi(Trẻ và Cao):

-Xét thuộc tính Hôn Nhân:

+ Gini cho node Độc Thân: 0.52 + 0.52 = 0.5

+ Gini cho node Đã Kết Hôn: 0.672 + 0.332 = 0.5578

* Trọng số Gini cho việc phân nhánh theo Hôn Nhân:

0.5 + 0.5578 = 0.5248

-Xét thuộc tính BĐS:

+ Gini cho node Nhà Thuê: 0.52 + 0.52 = 0.5

+ Gini cho node Nhà Sở Hữu: 0.672 + 0.332 = 0.5578

* Trọng số Gini cho việc phân nhánh theo BĐS:

0.5 + 0.5578 = 0.5248

Ta thấy, Gini Score cho Hôn Nhân bằngBĐS. Do đó việc phân nhánh Tuổi(trẻ và cao) sẽ dựa trên thuộc tính Hôn Nhân hoặc BĐS đều được