### Feature ĐẦU VÀO

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CHEST PAIN | GOOD BLOOD CIRCULATION | BLOCKED ARTERIES | WEIGHT | HEART DISEASE |
| NO | NO | NO | 125 | NO |
| YES | YES | YES | 180 | YES |
| YES | YES | NO | 210 | NO |
| YES | NO | YES | 167 | YES |

### BƯỚC 1 CREATE A NEW DATA SET BẰNG CÁCH RANDOM DATA SET CHO TRƯỚC ORIFINAL DATA(DATA SET CHO TRƯỚC)

### CÁI NEW DATA SET CÓ THỂ BỊ TRÙNG NHAU RƠI VÀO TRƯỜNG HỢP DUPLICATE(KHÔNG VẤN ĐỀ GÌ CẢ)

### NEW DATA SET

### (BOOSTRAPPED DATASET)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CHEST PAIN | GOOD BLOOD CIRCULATION | BLOCKED ARTERIES | WEIGHT | HEART DISEASE |
| YES | YES | YES | 180 | YES |
| NO | NO | NO | 125 | NO |
| YES | NO | YES | 167 | YES |
| YES | NO | YES | 167 | YES |

### BƯỚC 2: GENNERATE DECISION TREES FROM THE BOOTSRAPPED DATASET USING FREDEFINED CONDITIONS

### Từ 1 tập original dataset có thể gennerate ra nhiều newdata set từ những tập newdataset thì sẽ xây dựng nhiều cây khác nhau sử dụng predefined conditions

### 

TRADITIONAL DECISION TREE

### TRADITIONAL DECISION TREE chọn cả 4 cái feature (đối với decisiontree tính gini hoặc entropy tốt nhất trong 4 cái feature

### PRE-DEFINED CONDITIONS chọn số feature ít hơn feature đầu vào (còn random forest chỉ chọn 2 cái feature cho gini hoặc entropy tốt nhất xong rồi sử dụng các feature còn lại xây dựng cây

### 

PRE-DEFINED CONDITIONS

### DECISION TRÊ WITH PRE-DEFINED CONDITIONS

### Sau khi xây dựng được cây

### 

### Chọn feature tốt nhất làm node gốc

### 

### MÌNH SẼ XÂY DỰNG ĐƯỢC 1 CÁI CÂY DỰA TRÊN DATASET

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CHEST PAIN | GOOD BLOOD CIRCULATION | BLOCKED ARTERIES | WEIGHT | HEART DISEASE |
| YES | YES | YES | 180 | YES |
| NO | NO | NO | 125 | NO |
| YES | NO | YES | 167 | YES |
| YES | NO | YES | 167 | YES |

### 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CHEST PAIN | GOOD BLOOD CIRCULATION | BLOCKED ARTERIES | WEIGHT | HEART DISEASE |
| NO | NO | NO | 125 | NO |
| YES | YES | YES | 180 | YES |
| YES | YES | NO | 210 | NO |
| YES | NO | YES | 167 | YES |

### 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CHEST PAIN | GOOD BLOOD CIRCULATION | BLOCKED ARTERIES | WEIGHT | HEART DISEASE |
| NO | NO | NO | 125 | NO |
| YES | YES | YES | 180 | YES |
| YES | YES | NO | 210 | NO |
| YES | YES | NO | 210 | NO |

### 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CHEST PAIN | GOOD BLOOD CIRCULATION | BLOCKED ARTERIES | WEIGHT | HEART DISEASE |
| NO | NO | NO | 125 | NO |
| NO | NO | NO | 125 | NO |
| YES | YES | NO | 210 | NO |
| YES | NO | YES | 167 | YES |

### 

### KHI MÀ CÓ N cây randomforest thực hiện bằng classification

### 

### Sử dụng voting để đưa ra creadit dự đoán xem bệnh nhân có bệnh nhân không ví dụ có 20 cây có 13 cây đưa ra yes còn 7 cây đưa ra no thì quyết định bệnh nhân đó bị bệnh tim. Ngược lại thì bệnh nhân đó không bị bệnh tim

### Còn regression thì mình sẽ tính trung bình

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CHEST PAIN | GOOD BLOOD CIRCULATION | BLOCKED ARTERIES | WEIGHT | HEART DISEASE |
| NO | NO | NO | 125 | NO |
| YES | YES | YES | 180 | YES |
| YES | YES | NO | 210 | NO |
| YES | NO | YES | 167 | YES |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CHEST PAIN | GOOD BLOOD CIRCULATION | BLOCKED ARTERIES | WEIGHT | HEART DISEASE |
| YES | YES | YES | 180 | YES |
| NO | NO | NO | 125 | NO |
| YES | NO | YES | 167 | YES |
| YES | NO | YES | 167 | YES |
| CHEST PAIN | GOOD BLOOD CIRCULATION | BLOCKED ARTERIES | WEIGHT | HEART DISEASE |
| YES | YES | NO | 210 | NO |

### Khi random có sample không được select không nằm trong bootstrap dataset gọi là OUT-OF-BAG chúng ta sẽ sử dụng để testing đánh giá mô hình randomforest

### 

### 

### 

### Cuối cùng , có thể do lường được mức độ chính xác của random forest của chúng tôi bằng Out-Of-Bag chính xác bởi Random forest

### 

### Random forest nó có thể nâng cao thuật toán decision tree để feeling missing value

### Ví dụ:

### 

### Cách random forest feeling

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CHEST PAIN | GOOD BLOOD CIRCULATION | BLOCKED ARTERIES | WEIGHT | HEART DISEASE |
| NO | NO | NO | 125 | NO |
| YES | YES | YES | 180 | YES |
| YES | YES | NO | 210 | NO |
| YES | NO | N/A | N/A | NO |

### 

### Đầu tiền đoán một giá trị bất kì xong tinh chỉnh giá trị đó

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CHEST PAIN | GOOD BLOOD CIRCULATION | BLOCKED ARTERIES | WEIGHT | HEART DISEASE |
| NO | NO | NO | 125 | NO |
| YES | YES | YES | 180 | YES |
| YES | YES | NO | 210 | NO |
| YES | NO | NO | 167.5 | NO |

### Thấy giá trị nào xuất hiện nhiều hơn thì điền, còn giá trị liên tục thì điền trung bình.

### Sau khi có nhưng giá trị thì sẽ tinh chỉnh

### 

### Tiếp theo xây dựng một Proximity Matrix Of The First Tree

### 

### 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 |
| 1 |  |  |  |  |
| 2 |  |  |  |  |
| 3 |  |  |  | 1 |
| 4 |  |  | 1 |  |

### Đánh giá hai sample 3 and sample 4 có gần nhau hay không mình sẽ đánh giá bằng cách đưa vào decision stree có đưa ra cùng một quyết định hay không

### Proximity matrix này sẽ xây dựng một ma trận đánh giá sample khi mà vào N tree có đưa ra cùng một quyết định hay không

### 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 |
| 1 |  | 2 | 1 | 1 |
| 2 | 2 |  | 1 | 1 |
| 3 | 1 | 1 |  | 8 |
| 4 | 1 | 1 | 8 |  |

### Tiếp tục cho qua hết n cây đã generate(tạo ra) đếm xem với sample 3 với sample 4 đưa ra bao nhiêu quyết định ở bảng trên có 4 sample nên đưa ra ma trận 4x4 còn nếu nhiều sample hơn thì mình sẽ đưa ra ma trần cùng kích cở với những sample.

### Ví dụ sample 1 với sample 2 đưa ra cùng 2 quyết định sample 3 với 1 cùng có 1 quyết định sample 3 với sample 4 có cùng 8 quyết định

### Chúng ta có thấy ma trận này bằng nhau qua đường chéo chính đó gọi là proximy matrix để đánh giá các sample có closed(gần nhau hay không) qua N trees trong random forest có 10 cây mình sẽ normal line bằng cách chia đi 10 để về khoảng (0,1)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 |
| 1 |  | 0,2 | 0,1 | 0,1 |
| 2 | 0,2 |  | 0,1 | 0,1 |
| 3 | 0,1 | 0,1 |  | 0,8 |
| 4 | 0,1 | 0,1 | 0,8 |  |

### Bây giờ dựa vào matrix này mình sẽ refine lại các giá trị missing value này

### Bằng cách guess

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CHEST PAIN | GOOD BLOOD CIRCULATION | BLOCKED ARTERIES | WEIGHT | HEART DISEASE |
| NO | NO | NO | 125 | NO |
| YES | YES | YES | 180 | YES |
| YES | YES | NO | 210 | NO |
| YES | NO | N/A | N/A | NO |

### 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CHEST PAIN | GOOD BLOOD CIRCULATION | BLOCKED ARTERIES | WEIGHT | HEART DISEASE |
| NO | NO | NO | 125 | NO |
| YES | YES | YES | 180 | YES |
| YES | YES | NO | 210 | NO |
| YES | NO | NO | N/A | NO |

### Bây giờ chúng ta đến với sample 4 lúc nãy chúng ta tính trung bình (125\*1+180\*1+210\*1)/3=167.5

### Bây giờ chúng ta tìm các trọng số a b c

### 

### 

### Đấy là ứng dụng của random forest để chúng ta filling missing data