****HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



****

**BÀI TẬP LỚN**

**DỰ ĐOÁN NGUY CƠ MẮC BỆNH UNG THƯ VÚ**

**Nhóm lớp học: nhóm 9**

**Giảng viên giảng dạy: Gv. Vũ Hoài Nam**

|  |  |
| --- | --- |
| **Nguyễn Công Hiệp** | **B20DCCN239** |
| **Nguyễn Phú Tiến** | **B20DCCN596** |
| **Hoàng Quốc Việt** | **B20DCCN724** |

***Hà Nội – 2023***

1. **Tổng quan**

Chủ đề “Chuẩn đoán bệnh Ung thư vú”: Ung thư vú là loại ung thư phổ biến nhất ở phụ nữ trên toàn thế giới. Nó chiếm 25% tổng số ca ung thư và ảnh hưởng đến hơn 2,1 triệu người chỉ trong năm 2015. Ung thư vú bắt đầu khi các tế bào trong vùng vú bắt đầu phát triển không kiểm soát. Những tế bào này thường tạo thành khối u có thể được nhìn thấy thông qua tia X hoặc cảm nhận được dưới dạng cục u trong vùng vú.

Sử dụng các thư viện và framework như Sklearn, Numpy, Pandas và Mô hình học máy (Random forest, Descision tree Classification, Linear regression,… ) để so sánh hiệu quả

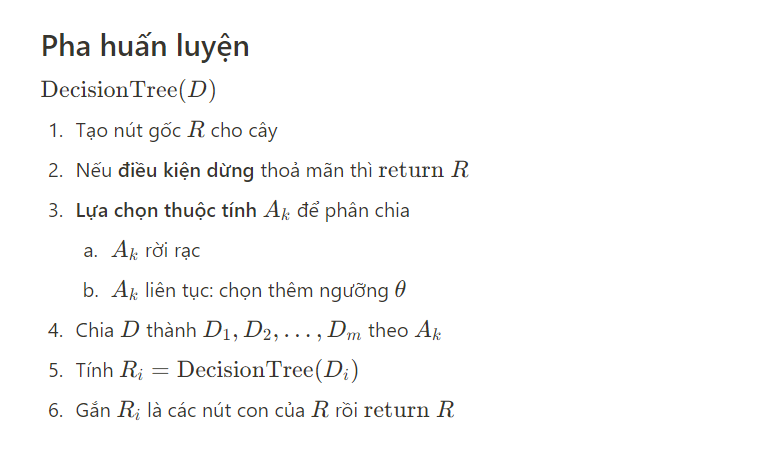
1. **Thuật Toán**

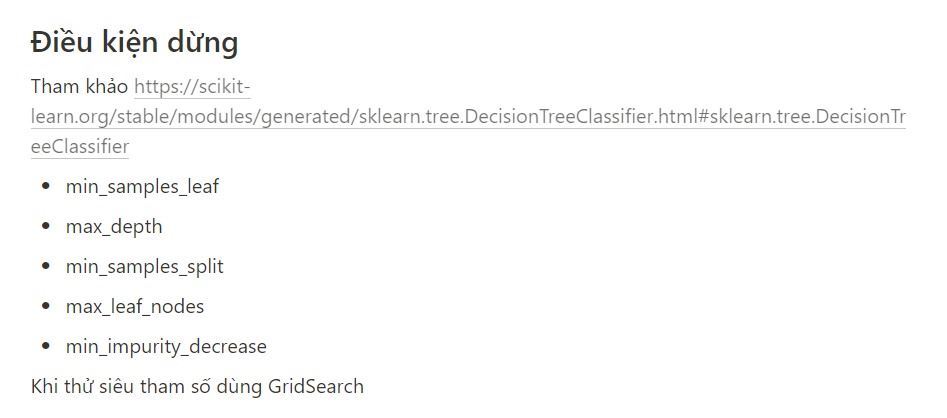
1.Giới thiệu thuật toán

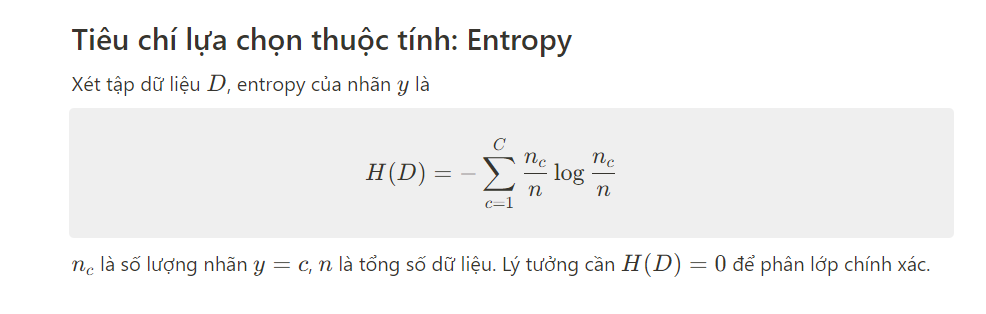
Giải thuật ID3 (gọi tắt là ID3) Được phát triển đồng thời bởi Quinlan trong AI và Breiman, Friedman, Olsen và Stone trong thống kê. ID3 là một giải thuật học đơn giản nhưng tỏ ra thành công trong nhiều lĩnh vực. ID3 là một giải thuật hay vì cách biểu diễn tri thức học được của nó, tiếp cận của nó trong việc quản lý tính phức tạp, heuristic của nó dùng cho việc chọn lựa các khái niệm ứng viên, và tiềm năng của nó đối với việc xử lý dữ liệu nhiễu.

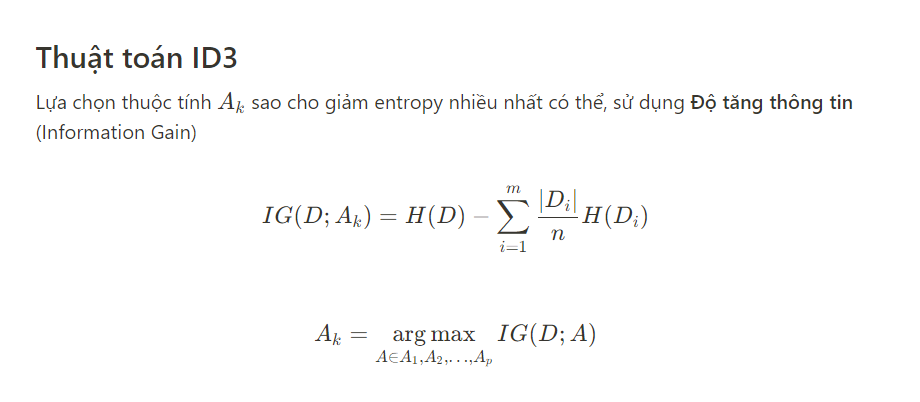
ID3 biểu diễn các khái niệm (concept) ở dạng các cây quyết định (decision tree). Biểu diễn này cho phép chúng ta xác định phân loại của một đối tượng bằng cách kiểm tra các giá trị của nó trên một số thuộc tính nào đó.

Như vậy, nhiệm vụ của giải thuật ID3 là học cây quyết định từ một tập các ví dụ rèn luyện (training example) hay còn gọi là dữ liệu rèn luyện (training data).

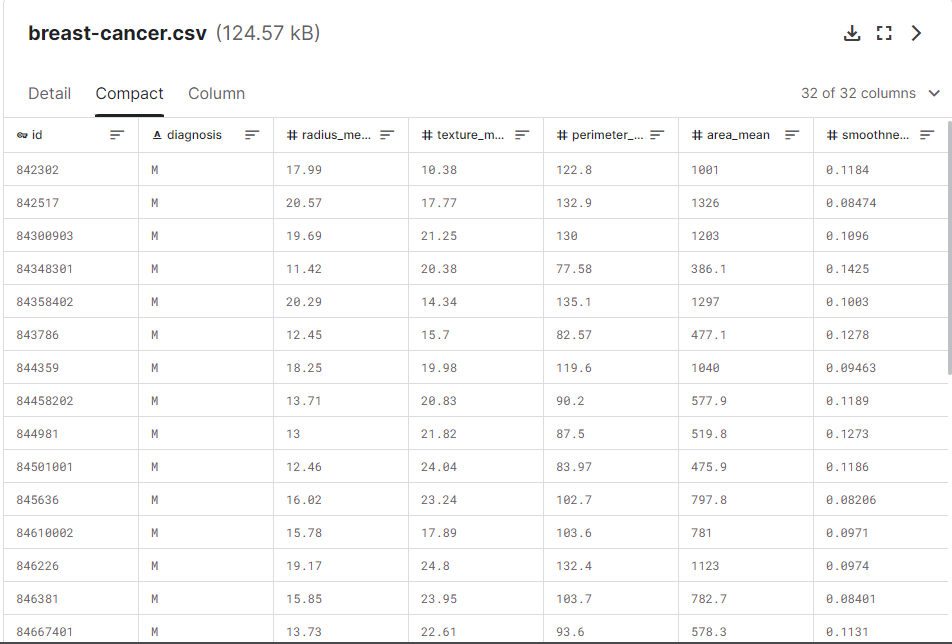
****

****

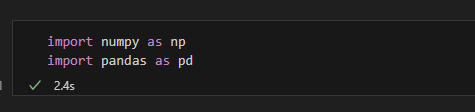
****

****

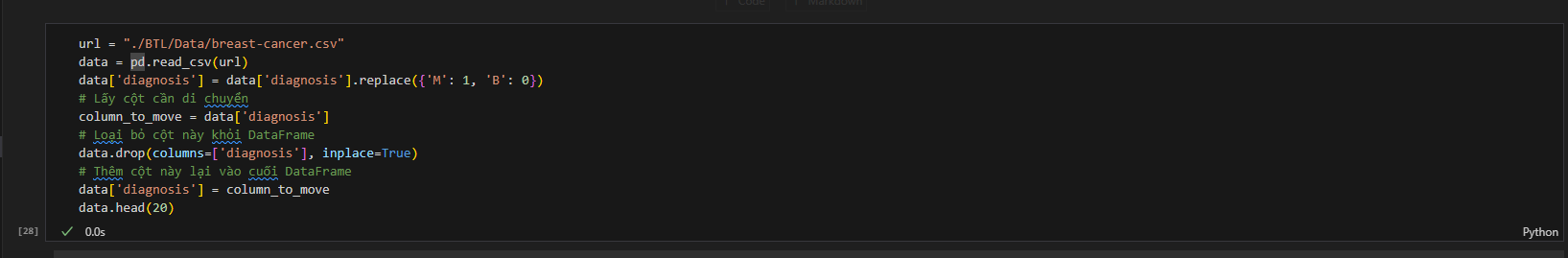
1. **Xây dựng chương trình**
2. **Data sử dụng**

****

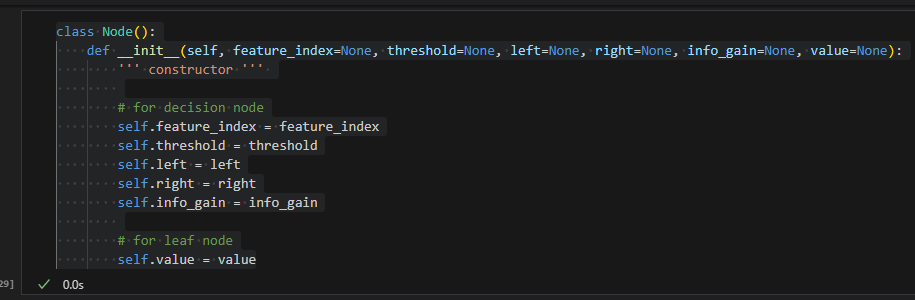
1. **Thư viện sử dụng**

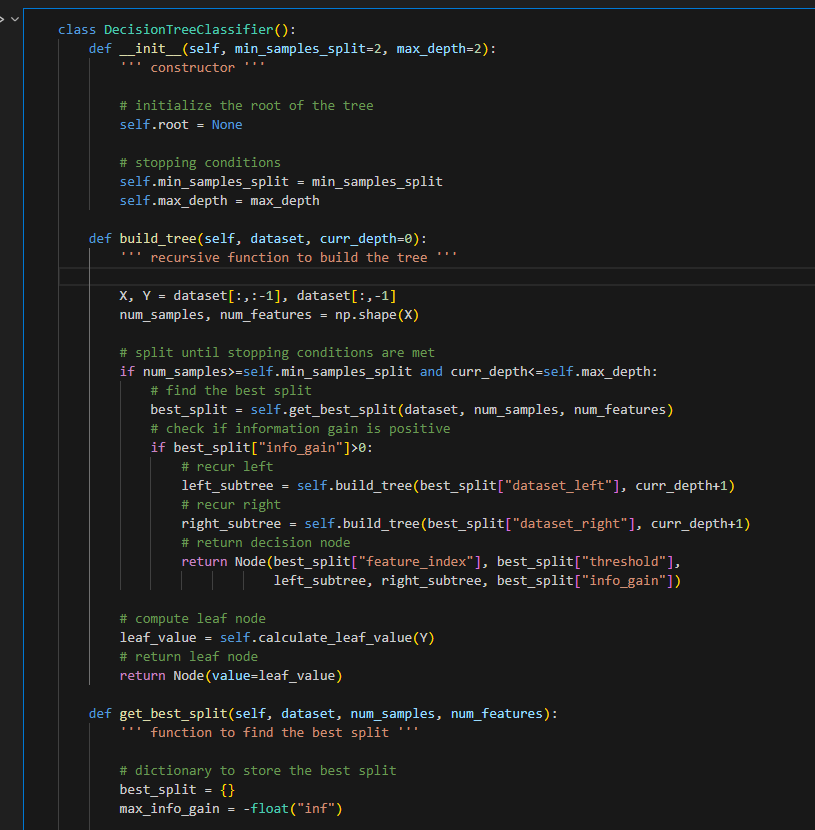
****

1. **Xử lí dữ liệu**

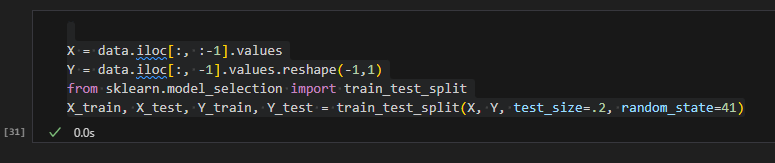
****

1. **Xây dựng mô hình Descision Tree**

****

****

1. **Chia tập dữ liệu**

****

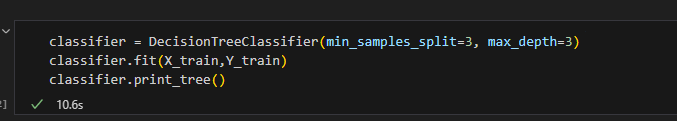
X\_train, X\_test: Tập huấn luyện và tập kiểm tra của các đặc trưng.

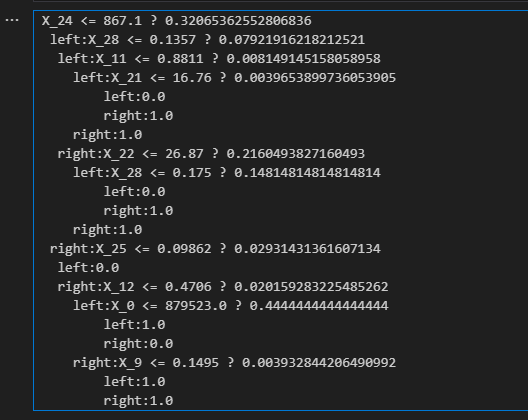
Y\_train, Y\_test: Tập huấn luyện và tập kiểm tra của biến phụ thuộc hoặc nhãn.

test\_size=0.2: Chỉ định rằng 20% của dữ liệu sẽ được sử dụng làm tập kiểm tra. Bạn có thể điều chỉnh giá trị này theo nhu cầu của mình.

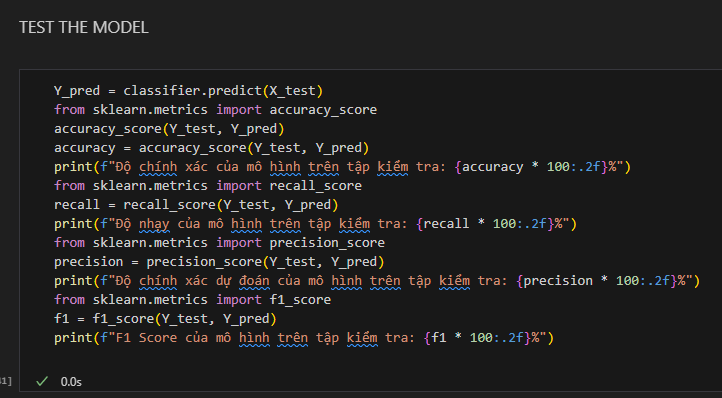
random\_state=41: Một số nguyên để đảm bảo rằng mỗi lần bạn chạy mã, việc chia tập dữ liệu sẽ luôn cho kết quả như nhau. Nếu bỏ qua tham số này, mỗi lần chạy có thể tạo ra một chia tập dữ liệu khác nhau.

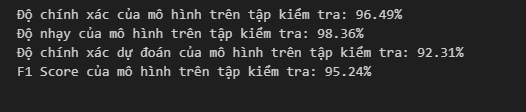
1. **Huấn luyện**

****

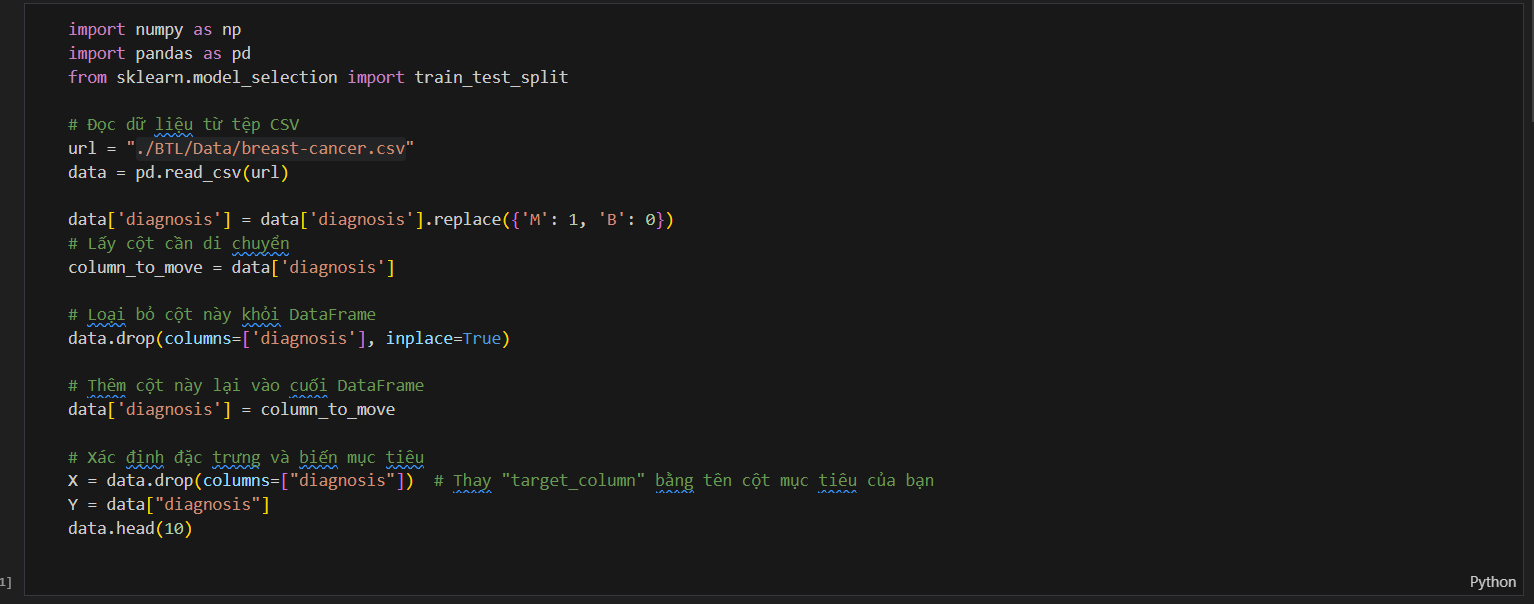
****

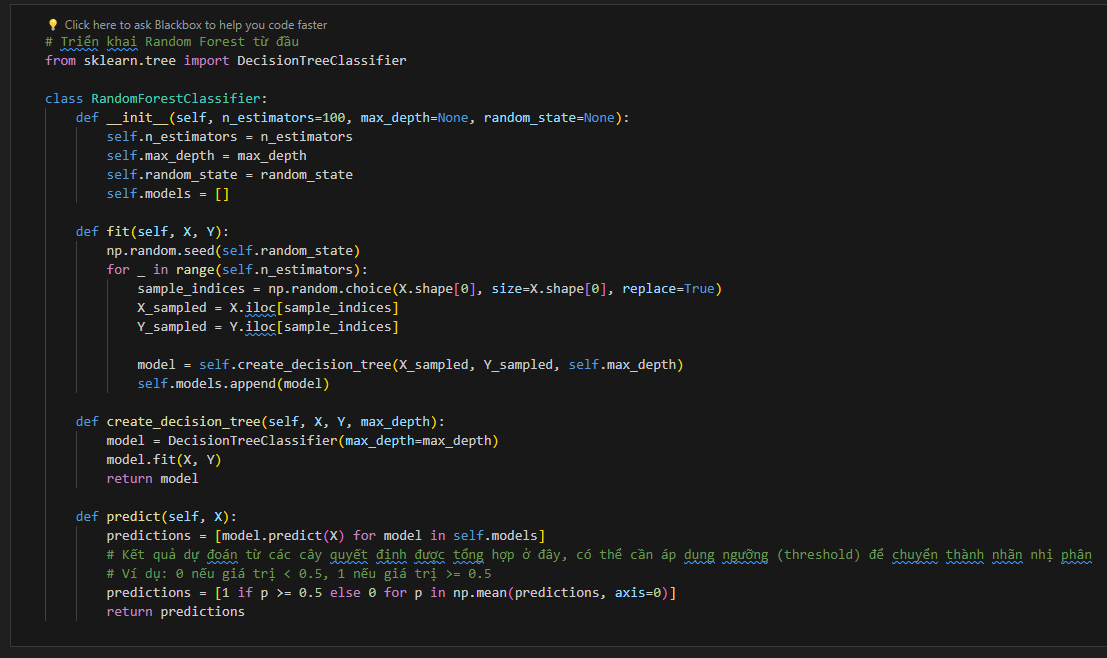
1. **Đánh giá mô hình**

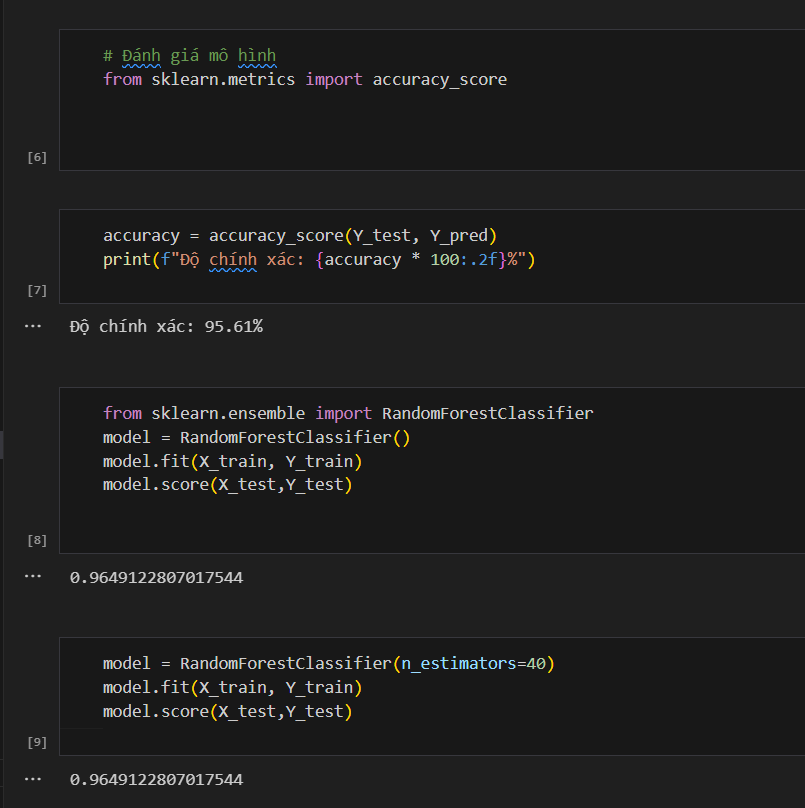
****

****

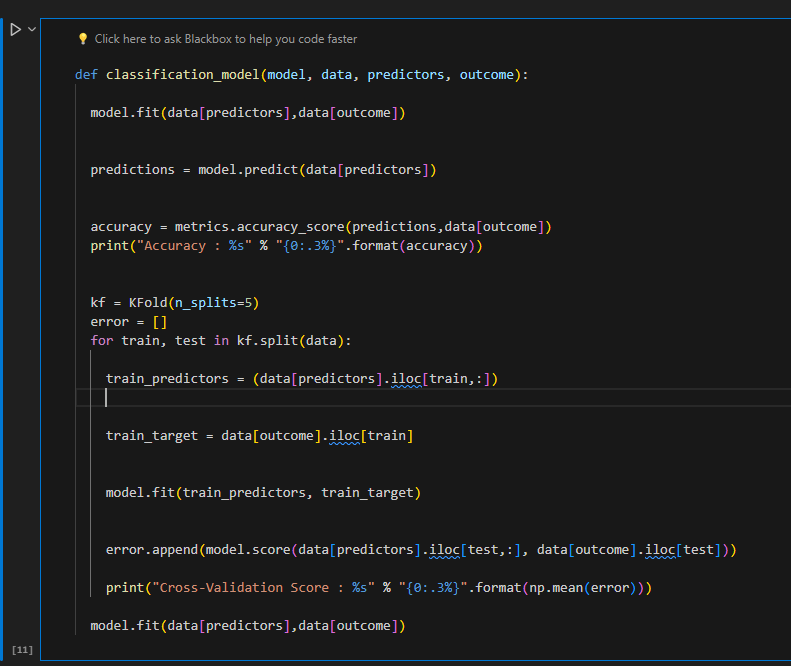
1. **Xây dựng mô hình Random Forest**

****

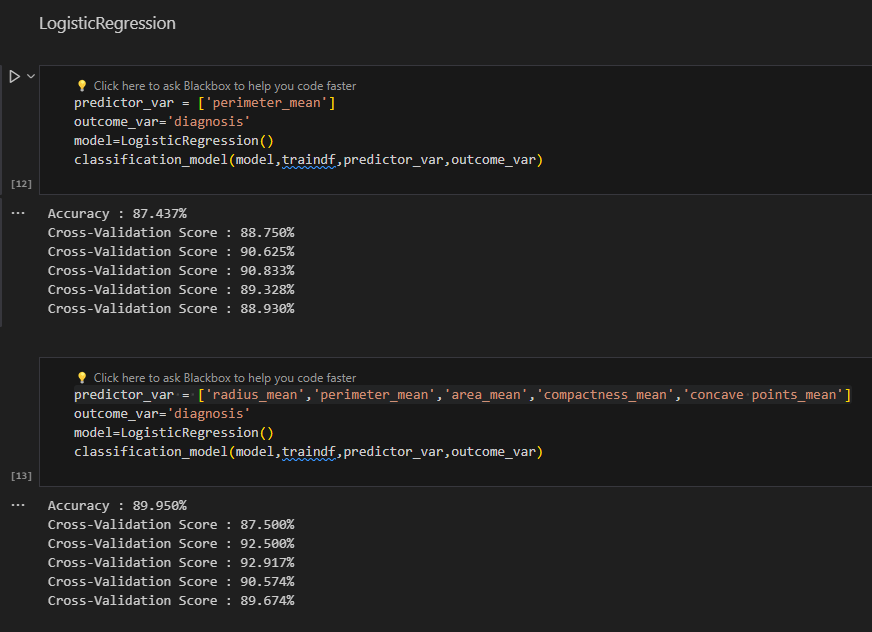
****

****

**11. Đánh giá so sánh 3 mô hình với các đặc trưng**

****

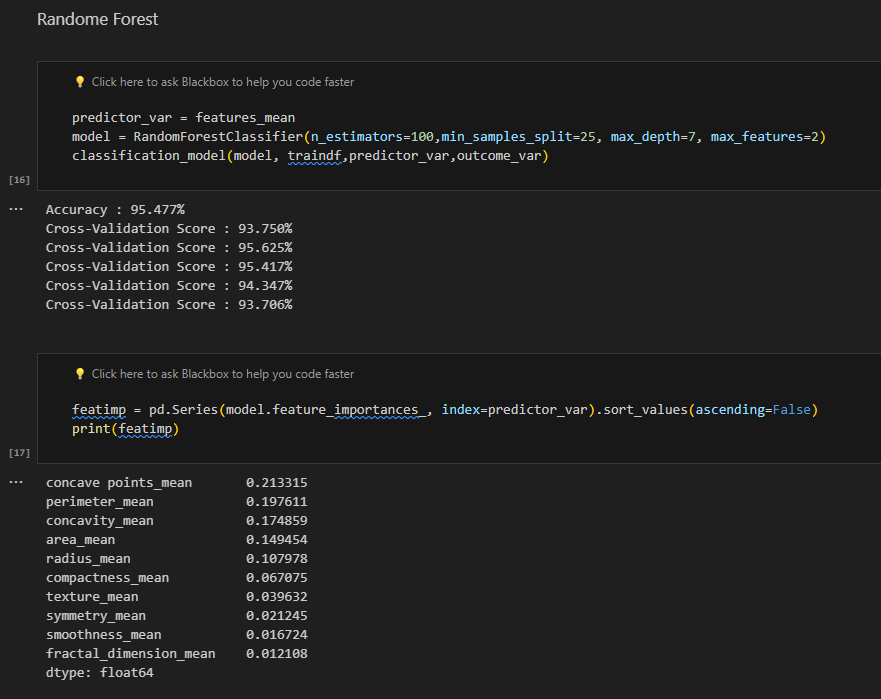
1. **LogisticRegression**

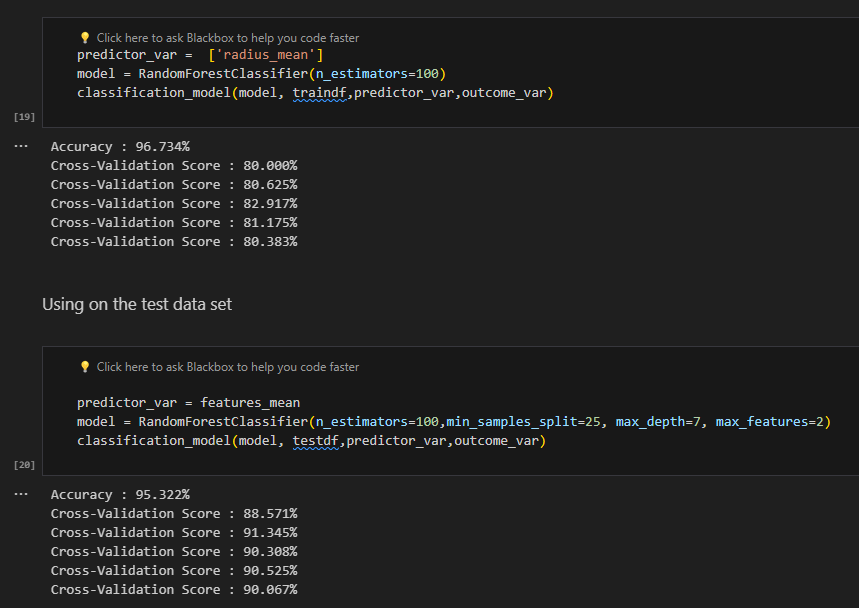
****

1. **Decision Tree model**

****

1. **Randome Forest**

****

****

1. **Kết luận**

* Mô hình Descision Tree đạt kết quả tốt
* Mô hình Descision Tree và Random Forest đều cho ra những chuẩn đoán có độ chính xác cao về nguy cơ mắc bệnh của bệnh ung thư vú là lành tính hay ác tính.
* Mô hình tốt nhất cho việc chẩn đoán ung thư vú, theo phân tích này, là Random Forest với các đặc trưng hàng đầu: 'concave points\_mean', 'area\_mean', 'radius\_mean', 'perimeter\_mean', 'concavity\_mean'. Mô hình này đạt đến độ chính xác dự đoán khoảng 95% và điểm chấp nhận được từ việc chia dữ liệu kiểm tra.
* Mô hình Random Forest có độ chính xác cao hơn và giảm nguy cơ overfitting.