Đại Học Quốc Gia Thành Phố Hồ Chí Minh

Đại Học Khoa Học Tự Nhiên

Khoa Công Nghệ Thông Tin

---

Logo

Description automatically generated with medium confidence

Project 2 – Triển khai mô hình phân loại bệnh ung thư vú

1. Thực hiện:
2. Môi trường:

* Ngôn ngữ: Python
* Công cụ: Jupyter Notebook cho xử lý dữ liệu và tkinter xây dựng UI cho application.
* File dữ liệu: data.csv và modified\_data.csv là phiên bản rút gọn của data.csv.
* Mô hình sử dụng: Logistic Regression.

1. Thành viên và phân công:

|  |  |
| --- | --- |
| MSSV | Tên |
| 19127592 | Lê Minh Trí |
| 19127587 | Trương Chí Toàn |
|  | Trần Quốc Việt |

|  |  |  |  |
| --- | --- | --- | --- |
| Công việc | Người thực hiện | Chi tiết | Hoàn thành |
| Chuẩn bị dữ liệu | Lê Minh Trí | Làm sạch dữ liệu, tiền xử lý, giảm thiểu trường dữ liệu không cần thiết | 100% |
| Thống kê chi tiết + UI | Trần Quốc Việt | Thống kê cơ bản các trường dữ liệu + xây dựng UI cơ bản | 100% |
| Trực quan dữ liệu + UI | Trương Chí Toàn | Trực quan các trường dữ liệu qua biểu đồ + xây dựng UI cơ bản | 100% |

1. Chuẩn bị dữ liệu:
   1. A black screen with white text

      Description automatically generated with low confidenceThu nhập dữ liệu

Graphical user interface, text

Description automatically generated

* 1. Khám phá dữ liệu

Graphical user interface, text

Description automatically generated

Bởi vì có 569 dòng và 33 cột, các trường đều thể hiện đầy đủ dữ liệu trừ cột cuối cùng "Unamed: 32". Ta sẽ drop cột này ở khâu tiền xử lý dữ liệu.

Ngoài ra, các trường dữ liệu dạng số đều là float64 nên không cần chuyển đổi dạng dữ liệu gì trong tập dữ liệu này.

* 1. Làm sạch dữ liệu

Để chắc chắn hơn ở các giá trị null, ta cần phải check xem các giá trị 0 ở các trường dữ liệu bởi vì đối với việc chẩn đoán bệnh và liên quan đến cơ thể con người,các giá trị cần được minh bạch.

Sau khi thực hiện kiểm tra các dòng dữ liệu có kết quả bằng 0, ta được kết quả sau…

Text

Description automatically generated

Hình ảnh chứng tỏ các giá trị 0 xuất hiện ở các cột liên quan đến concavity và có tất cả 13 dòng.

Với mỗi cột bất kì trong các trường dưới đây có giá trị bằng 0 thì các cột còn lại đều bằng 0.

A black screen with white text

Description automatically generated with low confidence

Graphical user interface, text, website

Description automatically generatedVậy, ta sẽ thực hiện làm sạch dữ liệu bằng cách bỏ các dòng này đi, cụ thể là 13 dòng như trên. Việc bỏ này là hợp lí vì số lượng các dòng này chỉ chiếm 2.28% trên tập dữ liệu tổng.

* 1. Tiền xử lý dữ liệu

Các dữ liệu đều là float và nằm trong khoảng cho phép nên ta không cần chuẩn hóa. Khi thực hiện mô hình hóa, các trường liên quan đến nhận dạng của từng bệnh nhân là không cần thiết vì ta chỉ cần quan tâm đến kết quả đạt được, ta có thể bỏ cột ID cũng như cột Unamed: 32 như đã nhắc ở phần khám phá dữ liệu.

Vẫn còn một vấn đề nữa, bởi vì đề yêu cầu ta nhập vào các giá trị đầu của các thuộc tính để bắt đầu chẩn đoán kết quả. Nhưng ta lại có tận 31 cột với các giá trị khác nhau để nhập, ta cần tìm sự tương quan giữa các trường với nhau để loại bỏ các trường không cần thiết và thu gọn tập dữ liệu.

Bằng cách áp dụng kỹ thuật heatmap, ta có thể trực quan hệ số tương quan giữa các biến với nhau:

Chart

Description automatically generated

Thông qua heatmap ở trên, ta có thể rút ra các nhận xét sau:

* + radius\_mean đối với area-mean và perimeter\_mean có sự tương quan cao lần lượt là 1 và 0.99 tức hai chỉ số này có liên quan mật thiết đến radius\_mean. Bên cạnh đó, area-mean và perimeter\_mean cũng có chỉ số tương quan là 0.99. Vậy ta có thể bỏ hai cột area-mean và perimeter\_mean.
  + Các chỉ số tương quan giữa compactness, concativity và concavepoints như đã nói Data Cleaning cũng có sự liên quan cao, cho nên ta cũng có thể áp dụng cách này để xóa các cột liên quan với nhau.
  + Tiếp theo, các chỉ số liên quan se cũng có sự tương quan cao khi radius\_es với area-mean và perimeter\_mean là 0.95 và 0.97. Tương tự, ta bỏ luôn hai cột này, giữ lại radius\_es.
  + Cuối cùng, ta có thể dễ dàng nhận ra các ô màu đỏ tập trung các chỉ số worst đối với chỉ số mean. Tức là các trường này cũng có chỉ số tương quan khá cao. Bởi vì các cột như area và perimeter đều theo radius nên điều tương tự sẽ xảy ra với các cột worst. Bên cạnh đó radius\_mean và radius\_worst cũng tương quan cao nên ta chỉ cần giữ lại radius\_mean là được.
* Chung quy lại, ta sẽ bỏ các cột liên quan đến perimeter và area(mean và es), các cột worst và cuối cùng là các cột compactness, concavity, và concave points.

Text

Description automatically generated

A black screen with white text

Description automatically generated with low confidenceKết quả cuối cùng đạt được, ta chỉ còn lại 1 trường dữ liệu thật sự cần thiết cho việc xây dựng và huấn luyện mô hình.

1. Huấn luyện mô hình:

Nhóm quyết định sử dụng Logistic Regression bởi vì mục đích của mô hình là dự đoán Yes hoặc No. Yes hoặc No ở đây chính là M(ác tính) hoặc B(lành tính) ở cột diagnosis, mô hình Logistic Regression đưa output trả ra về khoảng từ [0,1] từ hàm logit rất phù hợp cho mục đícht trên. Nếu kết quả output nào nhỏ hơn 0.5(điểm giữa của mô hình) thì được xem là M(ác tính) và ngược lại là B(lành tính)

* 1. Chia tập dữ liệu huấn luyện, kiểm thử

Ta có 4 tập dữ liệu chính:

* X\_train: các trường tham số cho mục đích huấn luyện
* X\_test: các trường tham số cho mục đích kiểm thử
* y\_train: cột diagnosis với các dòng cho mục đích huấn luyện
* y\_test: cột diagnosis với các dòng cho mục đích kiểm thử

Nhóm sử dụng thư viện statsmodels để thực hiện huấn luyện mô hình với hàm glm() (mô hình tuyến tính tổng quát, một loại mô hình bao gồm hồi quy logistic). Lý do sử dụng bởi vì nếu sử dụng trực tiếp LogisticRegression từ thư viện sklearn, các vấn đề overfitting có thể xảy ra. Glm thường trích xuất sự tuyến tính giữa các tham số đầu vào, tránh overfitting cho mô hình. Hơn nữa, Glm được xem là một mô hình tổng quát, việc áp dụng các mô hình khác ngoài Logistic cũng có thể xảy ra với sự tùy biến cao hơn.

Về vấn đề phân chia dữ liệu, ta sử dụng hàm train\_test\_split với test\_size = 0.3 và random\_state = 40.

Text

Description automatically generated

Ta sẽ chia tập dữ liệu theo tỷ lệ 7:3, 7 cho huấn luyện và 3 cho kiểm thử với số lượng khởi tạo là 40.

* 1. Thiết lập tinh chỉnh tham số

Ở phần chia tập huấn luyện, kiểm thử, random\_state ở một vài docs hay trên mạng thường để 42. Tuy nhiên khi đặt random\_state=40, nhóm lại phát hiện ra tỷ lệ chính xác lại cao hơn nên quyết định đặt là 40.

Text

Description automatically generatedKhi thiết lập tham số cho mô hình bằng statsmodels, ta không cần làm gì nhiều vì các tinh chỉnh sẽ được thực hiện sẵn. Cuối cùng, ta sẽ có bảng report về mô hình sau đây:

1. Báo cáo kết quả, so chính xác

Khi nhìn vào bảng report mô hình trên, ta có một vài vấn đề sau:

Text

Description automatically generatedTa cần tập trung vào giá trị P>|z|, theo lý thuyết thống kê, khi giá trị P càng lớn(cụ thể là vượt ngưỡng 0.05), độ tương quan giữa các trường với mô hình càng thấp hay gọi là giả thuyết vô hiệu(null-hypothesis). Việc bỏ các trường này sẽ không ảnh hưởng đến kết quả mô hình cũng như tỷ lệ chính xác. Ở đây ta cho ngưỡng là 0.1 để quan sát trước mô hình sau khi biến đổi sẽ xảy ra như thế nào.

Vậy các giá trị P sau khi thay đổi tập dữ liệu đã trở nên tốt hơn khi đều dưới 0.1 và rất nhiều biến nhỏ hơn 0.05 chứng tỏ mô hình đã tốt hơn trước rất nhiều.

Tiếp theo, ta trực quan mô hình cũng như tỷ lệ chính xác sau khi dự đoán kết quả thông quá tập test cho trước bằng classification\_report và confusion\_matrix.

Calendar

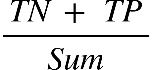
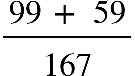
Description automatically generatedNhận xét:

Thông qua confusion\_matrix, ta có thể hiểu thêm về các thông số của dữ liệu, ta có:

* True positives(dương đúng) : 99
* True negatives(dương giả) : 59
* False positives(âm đúng) : 8
* False negatives(âm giả): 1

Ta có 100 bệnh nhân B(lành tính) và 67 bênh nhân M(ác tính)(sum=167). Với từng loại bệnh nhân ta có từng chỉ số mô hình khác nhau. Nhìn chung cả hai loại đều có tỷ lệ cao, hơn 90%(tuy tỷ lệ recall của M chỉ có 88% nhưng cũng được xem là cao).

Tóm gọn lại ở accuracy score, ta áp dụng công thức:

Accuracy score =  =  = 94.6%