**DỮ LIỆU BỆNH THẬN MÃN TÍNH**

Nhóm : 08

Thành Viên : Đặng Hưng Như Ý

GVHD : Trần Đình Toàn

MỤC LỤC

[1 Giới thiệu 3](#_Toc136468424)

[2 Hiểu biết về công việc 3](#_Toc136468425)

[3 Hiểu dữ liệu 4](#_Toc136468426)

[3.1 Trực quan khóa dữ liệu 4](#_Toc136468427)

[4 Chuẩn bị dữ liệu 7](#_Toc136468428)

[4.1 Làm sạch dữ liệu 7](#_Toc136468429)

[4.2 Chuyển đổi dữ liệu 8](#_Toc136468430)

[5 Lập mô hình và Đánh giá 10](#_Toc136468431)

[5.1 Chia tỷ lệ 10](#_Toc136468432)

[5.2 Tách 10](#_Toc136468433)

[5.3 Áp dụng các mô hình trên cơ sở 10](#_Toc136468434)

[5.4 Evaluation 15](#_Toc136468435)

[5.5 Lập mô hình với lựa chọn tính năng RFE 16](#_Toc136468436)

[5.6 Lập mô hình với lựa chọn tính năng CFS 17](#_Toc136468437)

[6 Triển khai 18](#_Toc136468438)

[7 Kết luận 18](#_Toc136468439)

# Giới thiệu

Ngày nay nhờ sự phát triển của máy tính và công nghệ, chúng ta có thể sử dụng các thuật toán học máy trong nhiều lĩnh vực khác nhau. Thật vậy, học máy là nghiên cứu sử dụng các thuật toán và dữ liệu cho phép máy tính thực hiện các tác vụ mà không cần hướng dẫn hoặc đầu vào từ người dùng.

Học tập hữu ích khi thiếu chuyên môn của con người, một lượng lớn dữ liệu cần xử lý hoặc ngay cả khi con người không thể giải thích chuyên môn của họ. Vì vậy, chúng tôi sử dụng máy học để phát triển các hệ thống có thể tự động khám phá kiến ​​thức mới từ cơ sở dữ liệu lớn, với khả năng bắt chước con người và thay thế một số nhiệm vụ đơn điệu.

Đối với dự án này, chúng tôi sẽ xử lý bộ dữ liệu Chẩn đoán bệnh thận mãn tính, trên đó chúng tôi sẽ đạt được khả năng tái tạo của các thí nghiệm khoa học được đề cập trong hai bài báo đã cho.

Để làm được điều đó, chúng tôi sẽ áp dụng phương pháp CRISP-DM, viết tắt của Quy trình tiêu chuẩn xuyên ngành để khai thác dữ liệu.

Trước hết chúng ta sẽ bắt đầu với Hiểu biết về Kinh doanh; các mục tiêu dự án và sự hiểu biết yêu cầu. Sau đó là hiểu biết về dữ liệu, tức là bước đầu thu thập và làm quen với dữ liệu. Tiếp theo là chuẩn bị dữ liệu để lựa chọn tính năng, chuyển đổi và làm sạch dữ liệu. Mô hình hóa, nơi chúng tôi áp dụng lựa chọn và ứng dụng kỹ thuật mô hình hóa. Việc đánh giá mô hình và cuối cùng là triển khai.

# Hiểu biết về công việc

Giai đoạn đầu tiên của quy trình CRISP-DM là hiểu những gì em muốn đạt được từ góc độ công việc. Trong trường hợp này, nghiên cứu này nhằm nâng cao chất lượng phân loại bệnh thận mạn tính bằng phương pháp chọn lọc đặc trưng và học tập đồng bộ.

Bệnh thận mãn tính (CKD) nằm trong số 20 nguyên nhân gây tử vong hàng đầu trên toàn thế giới và ảnh hưởng đến khoảng 10% dân số trưởng thành trên thế giới. CKD là tình trạng thận bị tổn thương và không thể lọc máu tốt như bình thường. Do đó, chất lỏng dư thừa và chất thải từ máu vẫn còn trong cơ thể và có thể gây ra các vấn đề sức khỏe khác, chẳng hạn như bệnh tim và đột quỵ.

Một số hậu quả sức khỏe của CKD bao gồm:

* Thiếu máu hoặc số lượng hồng cầu thấp
* Thiếu máu hoặc số lượng hồng cầu thấp
* Gia tăng các ca nhiễm trùng
* Nồng độ canxi thấp, nồng độ kali cao và nồng độ phốt pho cao trong máu gây thèm ăn hoặc ăn ít hoặc chất lượng cuộc sống thấp hơn
* huyết áp cao và thấp
* bệnh tiểu đường , tổn thương thần kinh và các vấn đề về xương

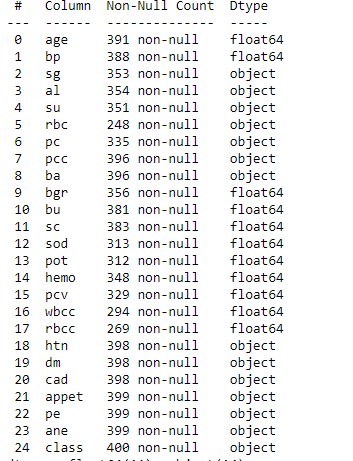
# Hiểu dữ liệu

## Trực quan khóa dữ liệu

Bộ dữ liệu CKD được thu thập từ 400 bệnh nhân.

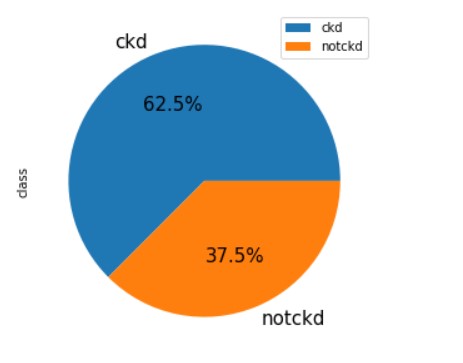
Hình 1 – Bộ dữ liệu gốc

Bộ dữ liệu bao gồm 24 tính năng được chia thành 11 tính năng số và 13 tính năng phân loại, ngoài các tính năng lớp, chẳng hạn như “ckd” và “notckd” để phân loại. Các tính năng bao gồm tuổi, huyết áp, trọng lượng riêng, albumin, đường, hồng cầu, tế bào mủ, khối tế bào mủ, vi khuẩn, đường huyết ngẫu nhiên, urê máu, creatinine huyết thanh, natri, kali, huyết sắc tố, thể tích tế bào đóng gói, tế bào bạch cầu số lượng, số lượng hồng cầu, tăng huyết áp, đái tháo đường, bệnh động mạch vành, thèm ăn, phù chân, và thiếu máu.



Hình 2 – Các tính năng

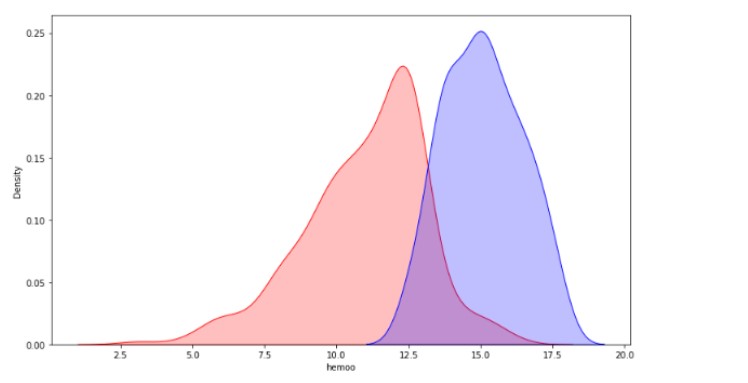
Lớp chẩn đoán chứa hai giá trị : ckd và notckd . Nó chứa 250 trường hợp thuộc loại “ckd” bằng 62,5% và 150 trường hợp thuộc loại “notckd” bằng 37,5%.



Hình 3 – Tỷ lệ người ckd và notckd

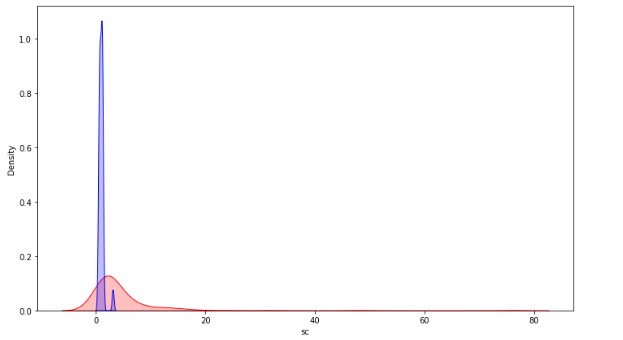
Để hiểu rõ hơn về tập dữ liệu, chúng tôi đã thực hiện các số liệu đại diện cho các tính năng khác nhau trong chức năng của lớp chẩn đoán.

**3.1.1 Mối tương quan giữa mục tiêu và các tính năng khác**

Trong hình dưới đây, chúng tôi lưu ý rằng đường cong màu đỏ đại diện cho bệnh nhân ckd, cho thấy Huyết sắc tố thấp có thể là dấu hiệu cho mức độ nghiêm trọng của bệnh thận

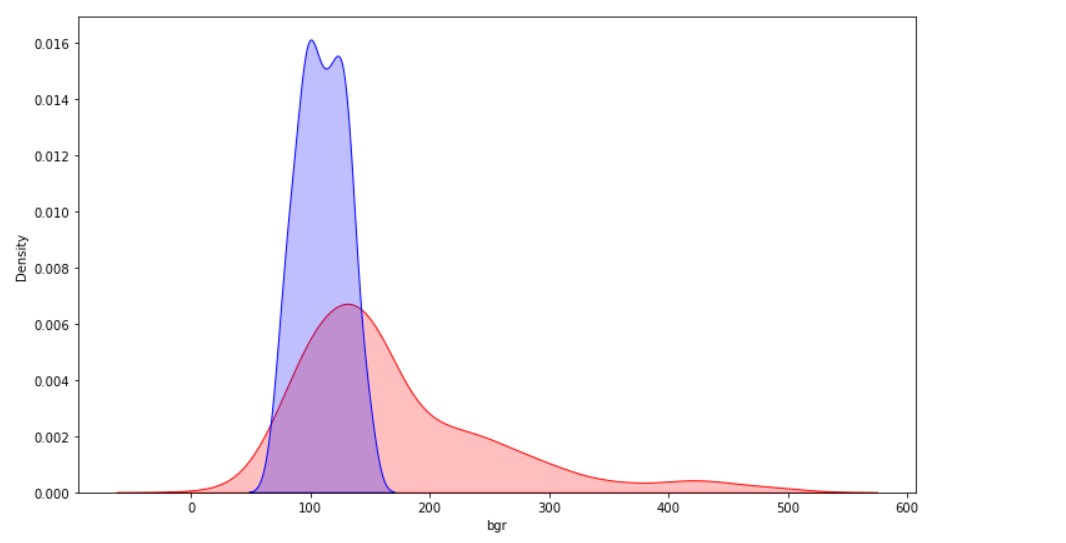
Hình 4 – Tương quan giữa các lớp và huyết sắc tố

Creatinine là một chất thải được tìm thấy trong máu trong các hoạt động cơ bắp. Thận tham gia vào việc loại bỏ chất thải này ra khỏi cơ thể và khi chức năng thận bị suy giảm, lượng creatinine còn lại trong máu sẽ cao.



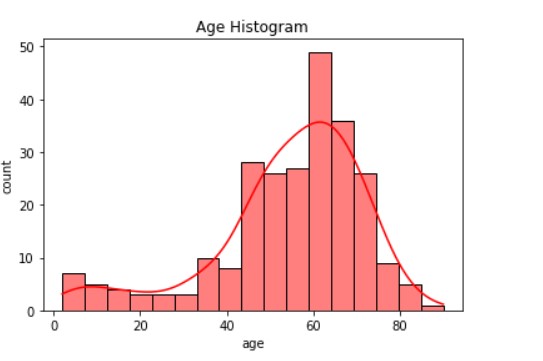
Hình 5 – Tương quan giữa các lớp và creatinine huyết thanh

Có nguy cơ hạ đường huyết ở bệnh nhân mắc bệnh thận mãn tính do chức năng thận suy giảm insulin và nếu bệnh nhân mắc bệnh tiểu đường, thuốc điều trị tiểu đường sẽ tồn tại trong hệ thống lâu hơn do giảm thanh thải thận.



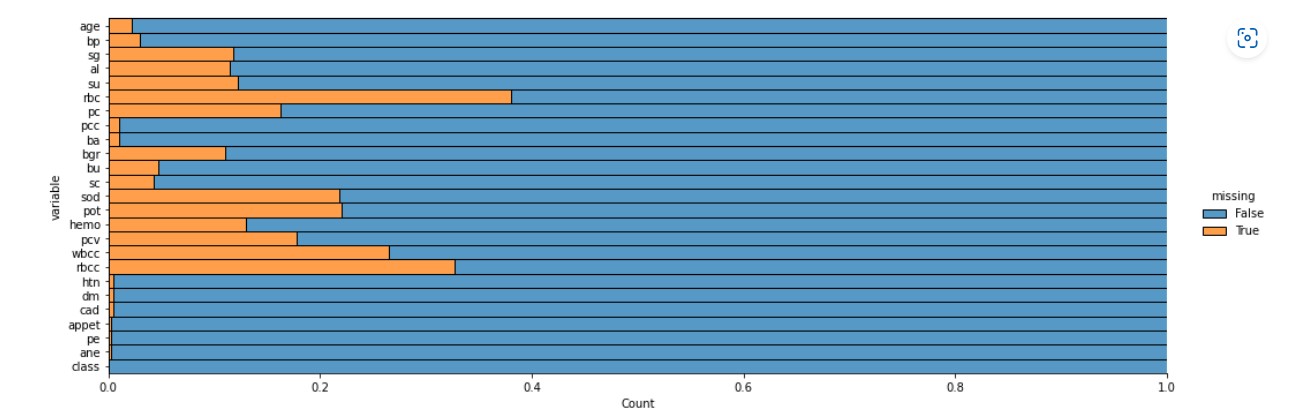
Hình 6 – Tương quan giữa các lớp và đường huyết huyết thanh

Theo số liệu hiện nay, CKD phổ biến hơn ở những người từ 60 đến 75 tuổi so với những người từ 45-60 tuổi hoặc 18-20 tuổi.



Hình 7 – Biểu đồ tuổi

**3.1.2 Trực quan hóa các giá trị còn thiếu**

Hình dưới đây cho thấy chúng ta có các giá trị bị thiếu, Hồng cầu có tỷ lệ giá trị bị thiếu cao nhất với 38% Và tỷ lệ thấp nhất là cảm giác thèm ăn và thiếu máu với 0,25%

Hình 8 – Phát hiện các giá trị bị thiếu

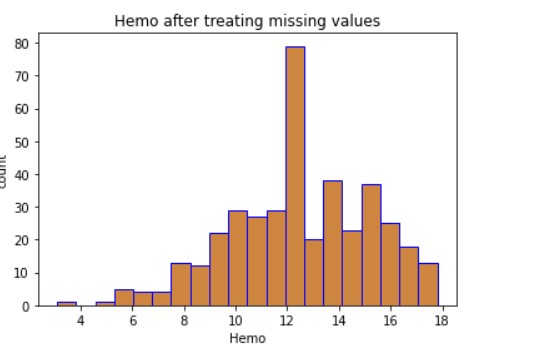
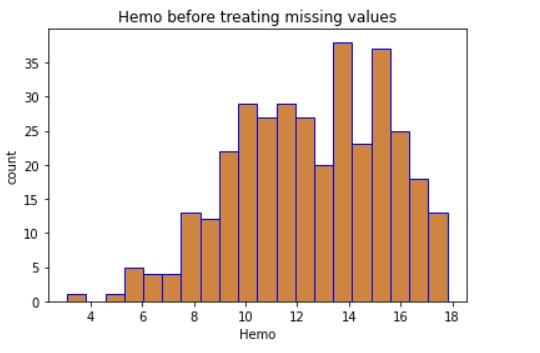
# Chuẩn bị dữ liệu

## Làm sạch dữ liệu

Làm sạch dữ liệu là quá trình sửa hoặc xóa dữ liệu bị hỏng, không chính xác hoặc không cần thiết khỏi khung dữ liệu.

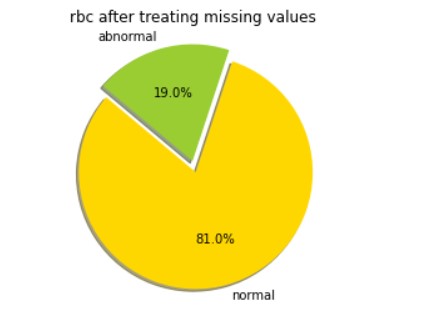
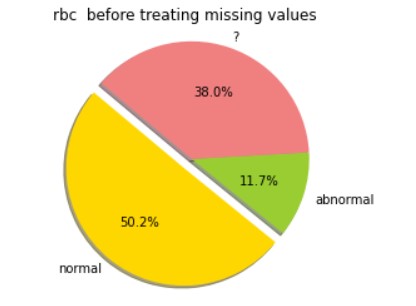
**4.1.1 Giá trị bị mất**

Để xử lý các giá trị thiếu số, chúng tôi chọn áp dụng phương pháp thống kê trung bình. Hai hình dưới đây cho thấy một ví dụ về đặc điểm số của huyết sắc tố trước và sau khi chuyển đổi



Hình 9 – Tính năng hemo trước và sau khi xử lý các giá trị bị thiếu

Để xử lý các giá trị thiếu danh nghĩa, chúng tôi chọn áp dụng các giá trị thường xuyên nhất. Hai hình dưới đây cho thấy một ví dụ về đặc trưng số rbc trước và sau khi chuyển đổi

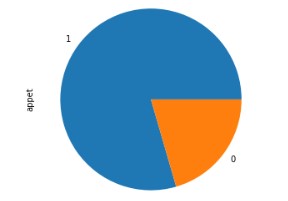
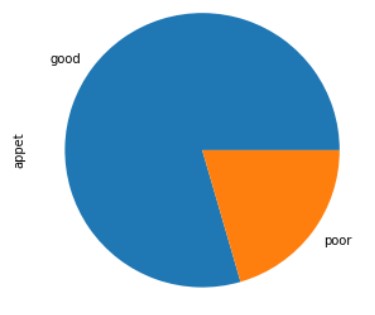


Hình 10 – Tính năng rbc trước và sau khi xử lý các giá trị bị thiếu

## Chuyển đổi dữ liệu

**4.2.1 Mã hóa**

Đối với Mã hóa , chúng tôi thường thay thế các giá trị danh nghĩa như ('bình thường', có', 'hiện tại', 'tốt') bằng 1 và tất cả dữ liệu phủ định ('bất thường','không' , 'không có mặt', 'kém') với 0 .



Hình 11 – Tính năng ứng dụng trước và sau khi mã hóa

**4.2.2 Lựa chọn tính năng**

Lựa chọn tính năng được sử dụng để giảm kích thước của tập dữ liệu và để chọn các tính năng phù hợp nhất được liên kết với CKD. Chúng tôi đã sử dụng RFE để lựa chọn tính năng cho việc học tập đồng bộ để cải thiện chẩn đoán CKD. Như đã đề cập trong hai bài báo này, Chúng tôi đã sử dụng mô hình tích hợp để chọn các tính năng tiêu biểu quan trọng nhất bằng cách sử dụng Thuật toán loại bỏ tính năng đệ quy (RFE)



Hình 12 – Đầu ra của thuật toán RFE

# Lập mô hình và Đánh giá

Việc lập mô hình bao gồm lựa chọn, định cấu hình và thử nghiệm các thuật toán khác nhau, cũng như quyết định trình tự của chúng để mang lại kết quả tốt nhất. Quá trình ban đầu là một quá trình mô tả tạo ra kiến ​​thức và giải thích tại sao mọi thứ lại xảy ra. Sau đó, nó trở thành dự đoán và giải thích điều gì sẽ xảy ra, và sau đó là quy định vì nó giúp tối ưu hóa các tình huống trong tương lai.

## Chia tỷ lệ

Chia tỷ lệ tính năng là quá trình chuẩn hóa phạm vi tính năng trong tập dữ liệu để các mô hình học máy diễn giải các tính năng này trên cùng một tỷ lệ.

## Tách

Bất cứ khi nào chúng tôi sử dụng mô hình học máy, chúng tôi không thể đào tạo mô hình đó trên một tập dữ liệu. Nếu chúng tôi đào tạo nó trên một tập dữ liệu thì chúng tôi sẽ không thể đánh giá hiệu suất của mô hình của mình. Vì lý do đó, chúng tôi chia dữ liệu nguồn của mình thành tập dữ liệu huấn luyện và thử nghiệm.

## Áp dụng các mô hình trên cơ sở

Trước hết, chúng tôi sẽ sử dụng tập dữ liệu mà không có bất kỳ lựa chọn tính năng nào.

**5.3.1 k-Nearest Neighbor (KNN)**

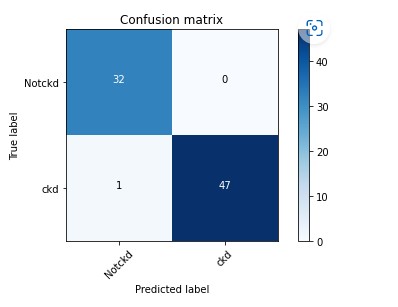
Thuật toán k-láng giềng gần nhất, còn được gọi là KNN hoặc k-NN, là một thuật toán phân loại học có giám sát, phi tham số, sử dụng khoảng cách gần để phân loại hoặc dự đoán về việc nhóm một điểm dữ liệu riêng lẻ. Mặc dù nó có thể được sử dụng cho các vấn đề hồi quy hoặc phân loại, nhưng nó thường được sử dụng như một thuật toán phân loại, dựa trên giả định rằng các điểm tương tự có thể được tìm thấy gần nhau.

* Tìm kiếm lưới

Để sử dụng đúng thuật toán KNN, chúng tôi đã áp dụng phương pháp Tìm kiếm lưới để xác định các tham số tốt nhất của mô hình. Tìm kiếm lưới thử các giá trị khác nhau rồi chọn giá trị mang lại điểm số cao nhất:

Số hàng xóm 3 và số liệu là euclidean

* Sau đó, chúng tôi áp dụng thuật toán sử dụng các tham số đó :



Hình 13- Ma trận nhầm lẫn

Kết quả của thuật toán dựa trên đầu ra của Ma trận nhầm lẫn như được đề cập dưới đây:

— Tích cực thực sự : Chúng tôi đã dự đoán 47 dương tính ( Ckd ) và đó là sự thật

— Âm tính thực sự : Chúng tôi đã dự đoán 32 âm tính (Không phải Ckd) và đó là sự thật

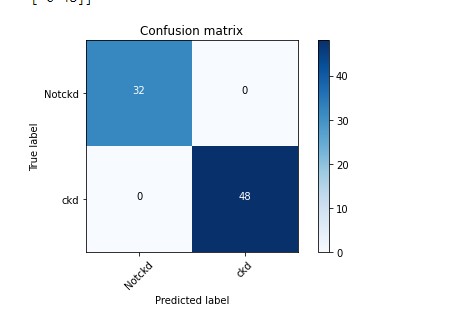
— Dương tính giả : chúng tôi đã dự đoán một dương tính giả

— Âm tính giả : Chúng tôi không dự đoán bất kỳ âm tính giả nào.

**5.3.2 Support Vector Machine ( SVM )**

Support Vector Machine (SVM) là một thuật toán học máy có giám sát được sử dụng cho cả phân loại và hồi quy. Mục tiêu của thuật toán SVM là tìm một siêu phẳng trong không gian N chiều phân loại rõ ràng các điểm dữ liệu.

Kết quả của thuật toán dựa trên đầu ra của Ma trận nhầm lẫn như được đề cập dưới đây:



Hình 14 – Ma trận nhầm lẫn

— Dương tính Đúng : Chúng tôi dự đoán 48 dương tính ( Ckd ) và điều đó đúng

— Âm tính Thật : Chúng tôi dự đoán 32 âm tính ( Không phải Ckd) và điều đó đúng

— Dương tính Sai : Chúng tôi không dự đoán bất kỳ âm tính giả nào.

— Âm tính giả : Chúng tôi không dự đoán bất kỳ âm tính giả nào.

**5.3.3 Naive Bayes :**

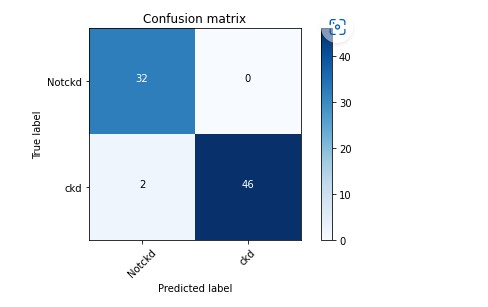
Trình phân loại Na¨ıve Bayes là một trong những thuật toán Phân loại đơn giản và hiệu quả nhất giúp xây dựng các mô hình máy học nhanh có thể đưa ra dự đoán nhanh. Nó là một bộ phân loại xác suất, có nghĩa là nó dự đoán trên cơ sở xác suất của một đối tượng. Kết quả của thuật toán dựa trên đầu ra của Ma trận nhầm lẫn như được đề cập dưới đây:

— Tích cực thực sự : Chúng tôi đã dự đoán 46 dương tính ( Ckd ) và đó là sự thật

— Âm tính thực sự : Chúng tôi đã dự đoán 32 âm tính (Không phải Ckd) và đó là sự thật

— Dương tính giả : chúng tôi đã dự đoán hai dương tính giả

— Âm tính giả : Chúng tôi đã dự đoán một dự đoán bất kỳ âm tính giả nào.



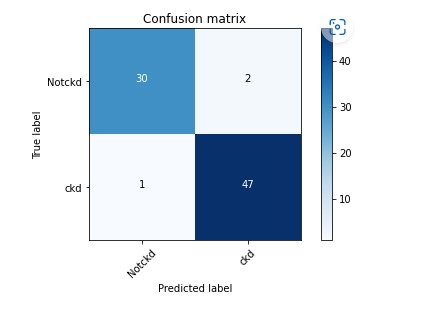
Hình 15 – Ma trận nhầm lẫn

**5.3.4 Decision Tree :**

Trình phân loại cây quyết định tạo mô hình phân loại bằng cách xây dựng cây quyết định. Mỗi nút trong cây chỉ định một phép thử trên một thuộc tính, mỗi nhánh đi xuống từ nút đó tương ứng với một trong các giá trị có thể có của thuộc tính đó

Chúng tôi đã xác định một tham số và sau đó áp dụng Tìm kiếm lưới để tìm các giá trị tốt nhất của tham số này là entropy

Ma trận nhầm lẫn:



Hình 16 – Ma trận nhầm lẫn

— Tích cực thực sự : Chúng tôi đã dự đoán 47 dương tính ( Ckd ) và đó là sự thật

— Âm tính thực sự : Chúng tôi đã dự đoán 30 âm tính (Không phải Ckd) và đó là sự thật

— Dương tính giả : chúng tôi đã dự đoán hai dương tính giả

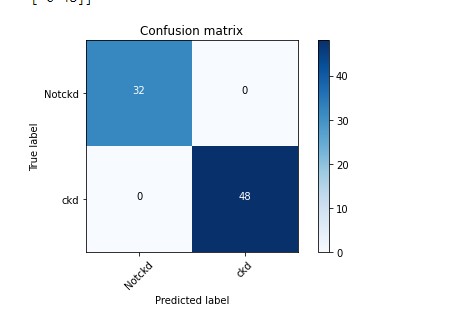
— Âm tính giả : Chúng tôi đã dự đoán một âm tính giả.

**5.3.5 Random Forest :**

Thuật toán rừng ngẫu nhiên được tạo thành từ một tập hợp các cây quyết định và mỗi cây trong tập hợp bao gồm một mẫu dữ liệu được lấy từ một tập huấn luyện có thay thế, được gọi là mẫu bootstrap.

Trước khi áp dụng thuật toán Rừng ngẫu nhiên, chúng tôi đã sử dụng phương pháp Tìm kiếm lưới để xác định các tham số tốt nhất:

Ma trận nhầm lẫn:



Hình 17 – Ma trận nhầm lẫn

— Tích cực thực sự : Chúng tôi dự đoán 48 kết quả tích cực thực sự.

— Âm tính thực sự : Chúng tôi dự đoán 32 âm tính thực sự.

— Dương tính giả : Chúng tôi không dự đoán bất kỳ dương tính giả nào.

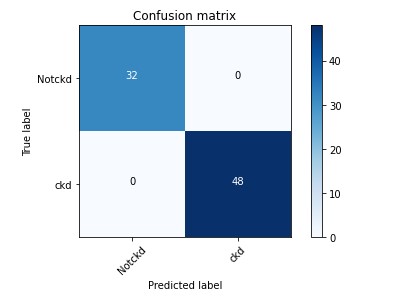
— Âm tính giả : chúng tôi không dự đoán bất kỳ âm tính giả nào.

**5.3.6 AdaBoost :**

Bộ phân loại AdaBoost là công cụ ước tính meta bắt đầu bằng cách khớp một bộ phân loại trên tập dữ liệu gốc, sau đó khớp các bản sao bổ sung của bộ phân loại trên cùng một tập dữ liệu nhưng trong đó trọng số của các phiên bản được phân loại không chính xác được điều chỉnh.

**5.3.7 Adaboost with Decision Tree :**

Cây quyết định được tăng cường bằng AdaBoost. Như chúng ta có thể nhận thấy từ ma trận ở trên, việc sử dụng mô hình này cho kết quả tương tự như SVM.

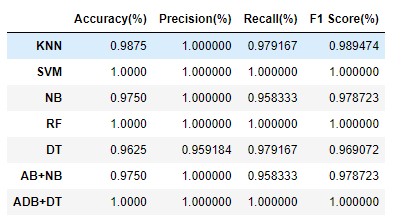


Hình 18 – Ma trận nhầm lẫn

## Evaluation

Sau khi áp dụng tất cả các mô hình được xử lý ở trên, đây là bảng tóm tắt tập hợp tất cả các kết quả.

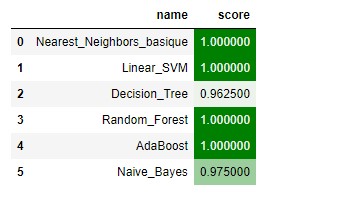
Dựa trên bảng, có thể lưu ý rằng Random Forest và SVM có độ chính xác cao nhất (100%), khả năng thu hồi cao nhất (100%), độ chính xác cao nhất (100%) và điểm F1 cao nhất (100%) so với tất cả các thuật toán khác . Độ chính xác của phương pháp 1 sử dụng bộ phân loại cơ sở khá cao, độ chính xác thấp nhất là 0,96 tỷ lệ chính xác với cây quyết định, cả Naive bayes và adaboost với tỷ lệ chính xác của Naive bayes là 0,958.



Hình 19 – Bảng tóm tắt các số liệu cơ bản

Từ các số liệu được ghi lại là tốt một cách phi thực tế, có một số cải tiến có thể được thực hiện để cải thiện hiệu suất phù hợp như lựa chọn tính năng, vì vậy chúng tôi sẽ áp dụng RFE và CFS.

hình bên dưới cho thấy điểm của từng bộ phân loại, như chúng ta có thể nhận thấy Knn, Adaboost, Random forest và SVM có điểm cao nhất.



Hình 20 – Bảng tóm tắt các số liệu cơ bản

## Lập mô hình với lựa chọn tính năng RFE

Bây giờ, chúng tôi sẽ sử dụng lựa chọn tính năng RFE để tăng hiệu suất của bộ phân loại cơ sở.

Loại bỏ tính năng đệ quy (RFE) làm giảm độ phức tạp của mô hình bằng cách loại bỏ từng tính năng một cho đến khi còn lại số lượng tính năng tối ưu. Đây là một trong những thuật toán lựa chọn tính năng phổ biến nhất do tính linh hoạt và dễ sử dụng của nó.

Các tính năng được xác định bởi RFE là:

albumin, đường, hồng cầu, tế bào mủ, kali creatinine huyết thanh, huyết sắc tố, số lượng hồng cầu, tăng huyết áp, đái tháo đường, thèm ăn và phù nề.

**k-Nearest Neighbor (KNN) :** Giống như chúng tôi đã làm khi áp dụng mô hình KNN trước đây, chúng tôi đã sử dụng Tìm kiếm lưới để xác định các tham số tốt nhất để sử dụng: Số lượng hàng xóm là 2 và số liệu là euclidean.

**Support vector machine (SVM) :** Các tham số tốt nhất cho mô hình sau khi Tìm kiếm lưới là: hạt nhân tuyến tính

**Decision tree (DT) :** Chúng tôi đã xác định một tham số và sau đó áp dụng Tìm kiếm lưới để tìm các giá trị tốt nhất của tham số này là entropy.

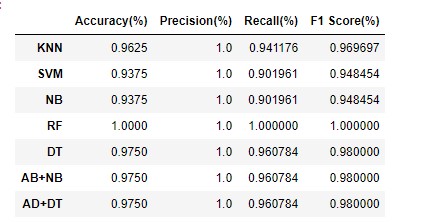
**Random forest :** Chúng tôi đã xác định một tham số và sau đó áp dụng Tìm kiếm lưới để tìm các giá trị tốt nhất của tham số này: 30 công cụ ước tính.

**5.5.1 Đánh giá với lựa chọn tính năng RFE**

Sau khi áp dụng tất cả các mô hình được xử lý ở trên, đây là bảng tóm tắt tập hợp tất cả các kết quả.

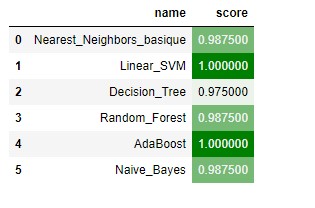
Dựa trên bảng, có thể lưu ý rằng Random Forest có độ chính xác cao nhất (100%), thu hồi cao nhất (100%), độ chính xác cao nhất (100%) và điểm F1 cao nhất (100%) trong tất cả các thuật toán khác.

Độ chính xác của phương pháp thứ nhất sử dụng bộ phân loại cơ sở RFE khá cao, độ chính xác thấp nhất là 0,93, tỷ lệ chính xác là SVM và NB, cả hai adaboost với Rừng ngẫu nhiên và quyết định tre đều có tỷ lệ chính xác là 0,97.



Hình 21 – Bảng tóm tắt các số liệu

hình bên dưới cho thấy điểm của từng bộ phân loại, vì chúng ta có thể nhận thấy SVM và Adaboost có điểm cao nhất.



Hình 22 – Bảng tóm tắt điểm số

## Lập mô hình với lựa chọn tính năng CFS

Lựa chọn tính năng dựa trên tương quan (CFS) phù hợp để áp dụng cho dữ liệu đa biến. CFS hoạt động bằng cách tính toán sự tương tác giữa các tính năng. CFS đánh giá một tập hợp con các tính năng có tính đến khả năng dự đoán của từng mức độ dư thừa giữa các tính năng và các tính năng đó.

Các đặc điểm được CFS xác định là: tăng huyết áp, Số lượng hồng cầu, thể tích tế bào đóng gói, khối tế bào mủ, creatinine huyết thanh, thiếu máu, đái tháo đường

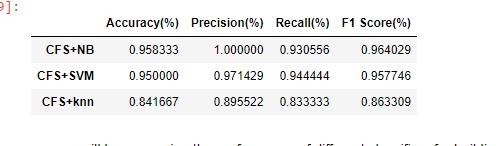
**k-Nearest Neighbor (KNN)** Giống như chúng tôi đã làm khi áp dụng mô hình KNN trước đây, chúng tôi đã sử dụng Tìm kiếm lưới để xác định các tham số tốt nhất để sử dụng: Số lượng hàng xóm là 1 và số liệu là euclidean

**Support vector machine (SVM)** Các thông số tốt nhất cho mô hình sau Grid

Tìm kiếm là : hạt nhân tuyến tính

**5.6.1 Đánh giá với lựa chọn tính năng CFS**

Sau khi áp dụng tất cả các mô hình được xử lý ở trên, đây là bảng tóm tắt tập hợp tất cả các kết quả. Dựa vào bảng, có thể thấy Naive bayes và SVM có độ chính xác cao nhất (95%). Độ chính xác của phương pháp thứ hai sử dụng bộ phân loại cơ sở RFE khá cao, tỷ lệ chính xác thấp nhất là 0,84 với KNN.



Hình 23 – Bảng tóm tắt các số liệu

hình bên dưới cho thấy điểm của từng bộ phân loại , vì chúng ta có thể nhận thấy SVM và Naive bayes có điểm cao nhất .



Bảng tóm tắt điểm số

# Triển khai

Bước cuối cùng trong phương pháp CRISP DM là triển khai trong đó việc khai thác dữ liệu được đền đáp. Trong giai đoạn cuối cùng này, những khám phá của em có thể xuất sắc đến đâu hay các mô hình của em phù hợp với dữ liệu hoàn hảo đến mức nào không quan trọng nếu em không thực sự sử dụng những điều đó để cải thiện cách thức kinh doanh của mình. Trong trường hợp của em cho dự án này, chúng tôi sẽ không xây dựng giai đoạn này.

# Kết luận

Báo cáo này thể hiện công việc của em trong việc khám phá và đào tạo bộ dữ liệu về bệnh thận mãn tính và thử các mô hình khác nhau.