**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH KHOA TOÁN - TIN HỌC**

**….****….**



BÁO CÁO KỸ THUẬT

Môn: Python cho Khoa Học Dữ Liệu

# 

# Đề tài: Dự đoán giá vàng

Giảng viên hướng dẫn: Hà Văn Thảo

Thành viên nhóm:

20280084 – Mai Chí Thanh 20280045 – Nguyễn Quốc Huy 20280109 – Mai Chí Trung

20280039 – Đặng Ngọc Hưng

**TP. Hồ Chí Minh, tháng 1 năm 2023**

# MỤC LỤC

[LỜI NÓI ĐẦU 3](#_bookmark0)

[TỔNG QUAN 4](#_bookmark1)

[DỰ ĐOÁN TUỔI THỌ SỬ DỤNG HỒI QUY TUYẾN TÍNH 5](#_bookmark2)

[Bước 1: Nhập các thư viện cần thiết 5](#_bookmark3)

[Bước 2: Tải bộ dữ liệu 5](#_bookmark4)

[Bước 3: Đánh giá thống kê 6](#_bookmark5)

[Bước 4: Trực quan hóa dữ liệu 8](#_bookmark6)

Bước 5: Features engineering and selection ……………………… 10

Bước 6: Model creation and training ……………………………… 12

[ĐÁNH GIÁ MÔ HÌNH ……………………………………………………. 1](#_bookmark7)3

Mô hình ANN ………………………………………………………. 13

**Mô hình** **DecisionTreeRegressor ………………………………......** 15

**Mô hình SVR ………………………………………………………..** 17

[KẾT LUẬN 11](#_bookmark8)

[TÀI LIỆU THAM KHẢO 12](#_bookmark9)

# LỜI NÓI ĐẦU

Trong lịch sử, vàng đã được sử dụng như một hình thức tiền tệ ở nhiều nơi trên thế giới bao gồm Việt Nam. Hiện tại, kim loại quý như vàng được giữ tại ngân hàng trung ương của tất cả các quốc gia để đảm bảo thanh toán nợ nước ngoài, đồng thời kiểm soát lạm phát, phản ánh sức mạnh tài chính của đất nước. Gần đây, các nền kinh tế mới nổi trên thế giới, chẳng hạn như Trung Quốc, Nga và Ấn Độ là những nước mua vàng lớn, trong khi Hoa Kỳ, Nam Phi và Úc là những quốc gia bán vàng nhiều nhất.

Dự báo tăng và giảm tỷ giá vàng hàng ngày, có thể giúp các nhà đầu tư quyết định khi nào nên mua (hoặc bán) vàng. Nhưng giá vàng còn phụ thuộc vào nhiều yếu tố như giá các kim loại quý khác, giá của dầu thô, giao dịch chứng khoán, giá trái phiếu, tỷ giá hối đoái, …

Trong dự án này sẽ dự báo tỷ giá vàng bằng cách sử dụng bộ tính năng toàn diện nhất và sẽ áp dụng các thuật toán khác nhau để dự báo và so sánh kết quả của chúng.

# TỔNG QUAN

Thách thức trong dự án này là dự đoán chính xác giá đóng cửa được điều chỉnh trong tương lai. Vấn đề là một vấn đề hồi quy, bởi vì giá trị đầu ra là giá đóng cửa được điều chỉnh trong dự án này là giá trị liên tục. Nhiều nghiên cứu khác nhau đã được các nhà nghiên cứu thực hiện để dự báo tỷ giá vàng bằng cách sử dụng các máy học khác nhau các thuật toán với mức độ thành công khác nhau nhưng cho đến gần đây khả năng xây dựng các mô hình này đã được giới hạn trong giới học thuật. Giờ đây, với các thư viện như Scikit-learning, bất kỳ ai cũng có thể xây dựng các mô hình dự đoán mạnh mẽ. Đối với dự án này, tôi sẽ sử dụng các mô hình học máy tuyến tính, tập hợp và tăng cường khác nhau để dự đoán giá đóng cửa đã điều chỉnh của Nasdaq sử dụng bộ dữ liệu về giá vàng theo thời gian thực tính bằng USD từ năm 2012 đến năm 2022

## DỰ ĐOÁN GIÁ VÀNG SỬ DỤNG HỒI QUY TUYẾN TÍNH

### BNGFHGHGHGBước 1: Nhập các thư viện cần thiết

Sử dụng Pandas để nhập dữ liệu, Matplotlib và Seaborn để trực quan hóa dữ liệu, sklearn cho các thuật toán, tran\_test\_split để chia tập dữ liệu trong tập kiểm tra và traing set, accuracy\_score để đánh giá mô hình.

Text

Description automatically generated

### DSFDWFFBước 2: Tải bộ dữ liệu

Lấy dữ liệu từ trang:

<https://www.nasdaq.com/marketactivity/commodities/gc%3Acmx>

Text

Description automatically generated

Bộ dữ liệu này bao gồm giá vàng theo thời gian thực tính bằng USD từ năm 2012 đến năm 2022.  
 Date - Ngày ghi giá  
 Close - Ngày đóng giá vàng tính bằng USD  
 Volume - Tổng lượng mua và bán của hàng hoá vàng  
 Open - Giá mở của vàng vào ngày cụ thể đó  
 High - Giá vàng cao nhất của ngày cụ thể đó  
 Low - Giá vàng thấp nhất vào ngày cụ thể đó.

Table

Description automatically generated

### 

### 

### Đọc dữ liệu vàng hàng ngày trong 10 năm qua và lưu trữ nó trong df. Xóa các cột không liên quan

### Chart, scatter chart Description automatically generated

### Bước 3: Đánh giá thống kê thông tin

### Graphical user interface, text Description automatically generated

### Các thuộc tính có kiểu dữ liệu phù hợp để phân tích, cột Date đã có kiểu Datetime

### Table Description automatically generated

### Kích thước của dữ liệu

### 

Dữ liệu có missing value nên phải xử lý lại dữ liệu

Graphical user interface, text, application

Description automatically generated

Sau khi xử lý missing value chúng ta loại bỏ đi 39 hàng trong bộ dữ liệu.

Do 39 dòng nhỏ so với số lượng hơn 2000 dòng trong bộ dữ liệu nên sau khi loại bỏ 39 dòng, bộ dữ liệu vẫn còn đủ tốt để phân tích và train model.

**Bước 4: Trực quan hóa dữ liệu**

A picture containing company name

Description automatically generated Sử dụng hàm pairplot, heatmap để tìm kiếm các biến có tương quan mạnh với giá vàng

A picture containing shape

Description automatically generatedChart

Description automatically generated

### 

Nhìn vào biểu đồ trên ta thấy các biến Close/Last, Open, High, Low có tương quan với nhau

Ngoài ra có thể sử dụng heatmap để tìm tương quan giữa các biến

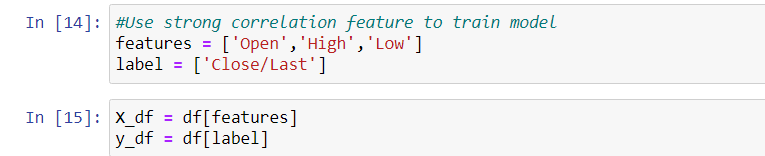
Bar chart

Description automatically generated with medium confidence

Qua biểu đồ trên cho thấy các biến Close/Last, Open, High, Low có tương quan mạnh với nhau

**Bước 5: Features engineering and selection**

Sử dụng tính năng tương quan để train model

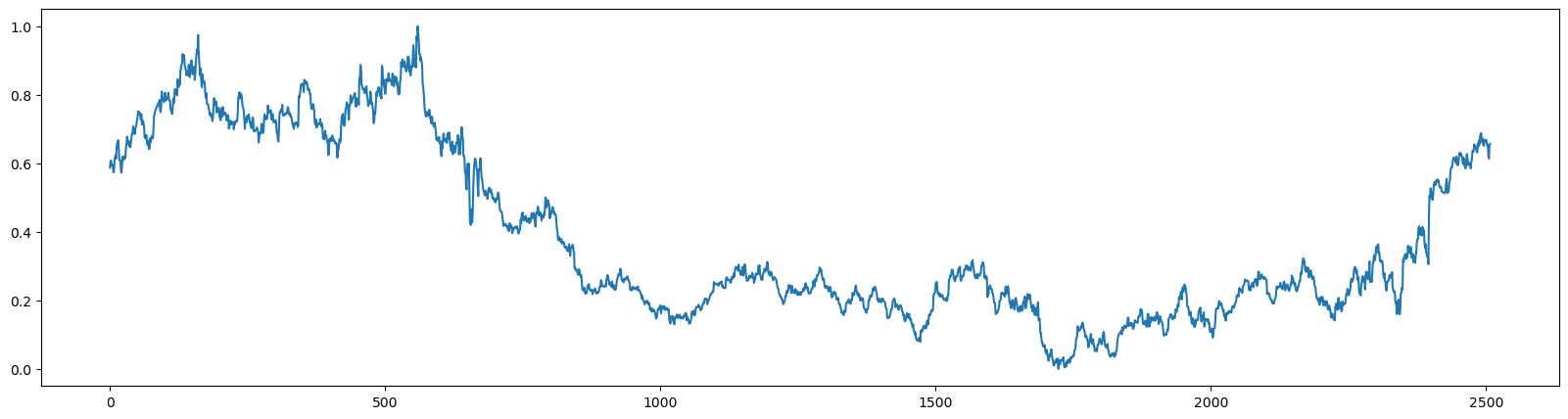
****

Text

Description automatically generated Sử dụng MinMaxScaler của sklearn để chuẩn hóa lại dữ liệu trong khoảng từ 0 đến 1

Text

Description automatically generated Phía dưới là dữ liệu sau khi biến đổi:

Chart, scatter chart

Description automatically generatedText

Description automatically generated So sánh dữ liệu trước và sau khi biến đổi

## Bước 6: Model creation and training

## Tiếp theo tách dữ liệu để train model theo kích thước 8:2

## Graphical user interface, text, application, email Description automatically generated

## Chart, line chart, scatter chart Description automatically generatedChart, line chart, scatter chart Description automatically generated

**ĐÁNH GIÁ MÔ HÌNH**

**Graphical user interface, text

Description automatically generated**

**Mô hình ANN:**

3 hidden layers

40 Neuron node on each layer

Relu activation function

Graphical user interface, text, email

Description automatically generated

Mô hình sẽ biên dịch với:

Chức năng tối ưu hóa Adam

Với hàm mất lỗi bình phương trung bình

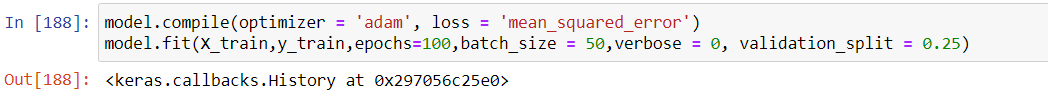
Tiến trình training với:

100 Epochs

50 rows of batch size

Coross validation 20% of training set

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application

Description automatically generated

Chart, histogram

Description automatically generated

**DecisionTreeRegressor:**

Graphical user interface, text, application, Word

Description automatically generated Sử dụng mô hình DecisionTree Regressor để dự đoán với tham số mặc định

Graphical user interface, text, application

Description automatically generated

Chart

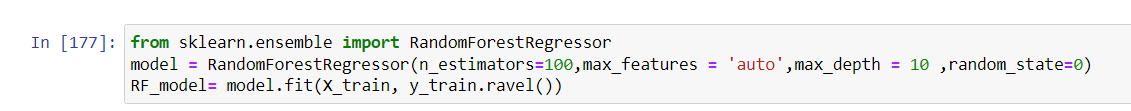
Description automatically generated

Từ kết quả trên, chúng ta có thế thấy rằng hiệu suất của mô hình thật sự tốt trên bộ xác thực, mô hình phần nào dự đoán giá tốt.

**RandomForest:**

Random Forest là một thuật toán học có giám sát. Nó tạo ra một khu rừng và biến nó thành ngẫu nhiên theo cách nào đó. Các rừng mà nó xây dựng, là một tập hợp các Cây quyết định, phần lớn thời gian được huấn luyện bằng phương pháp “đóng gói”. Các ý tưởng chung của phương pháp đóng gói là sự kết hợp của các mô hình học tập sẽ làm tăng kết quả tổng thể.Các khu rừng quyết định ngẫu nhiên sửa lỗi cho các cây quyết định có thói quen khớp quá mức với tập huấn luyện của chúng.Tôi đã sử dụng hai tham số n\_estimators=50 (giá trị mặc định =10), số lượng cây trong rừng. và random\_state=0, random\_state là hạt giống được sử dụng bởi trình tạo số ngẫu nhiên.

Trong phần này sẽ chỉnh 3 thông số của RamdomForest là n\_estimators, max\_features, max\_depth



Graphical user interface, text, application, email

Description automatically generated

Sau khi áp dụng random forest với n\_estimators = 100, đạt được kết quả như sau:

Graphical user interface, chart

Description automatically generated

**SVR:**

Mô hình được tạo ra bởi Support Vector Regression chỉ phụ thuộc vào một tập con của dữ liệu huấn luyện, bởi vì hàm chi phí để xây dựng mô hình bỏ qua mọi dữ liệu huấn luyện gần với dự đoán của mô hình.

Sử dụng SVR với (kernel=’linear’)

Graphical user interface, text, application

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Chart, histogram

Description automatically generated