**Report on the Mini Project**

Topic: Diamond Price Prediction

Algorithms: Linear Regression, Decision Trees, Random Forests, Neural Network

Team 10

Nguyễn Hữu Tiến 20205033

Nguyễn Hoàng Hải 2020xxxx

Kiều Doãn Trung 20205037

Course: Machine Learning – 2022.2

26/06/2022

1. Problem Description

Bộ dataset bao gồm giá và các thuộc tính khác của khoảng 54.000 viên kim cương, các thuộc tính này có thể ảnh hưởng đến giá của viên kim cương.

Bộ [diamonds dataset](https://www.kaggle.com/datasets/shivam2503/diamonds) có 53940 bản ghi và 10 thuộc tính bao gồm cả thuộc tính ở dạng số và ở dạng được phân loại. Nhiệm vụ của chúng tôi là dự đoán ra giá của viên kim cương dựa trên các thuộc tính liên quan này.

1. Exploratory Data Analysis

Trước tiên chúng tôi xem qua bộ dữ liệu để hiểu hơn về dữ liệu mà chúng tôi đang xử lý. Phân tích dữ liệu được để ở file notebook có tên exploratory.ipynb

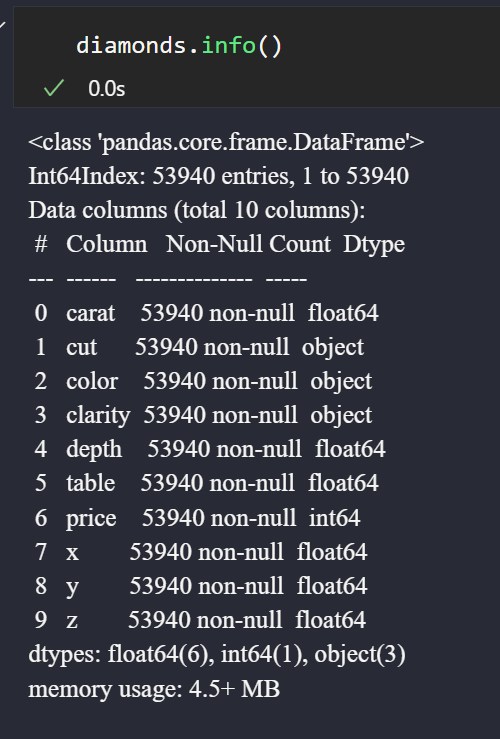
1. Data collection

Bộ dữ liệu có định dạng file csv được lấy tại trang web: <https://www.kaggle.com/datasets/shivam2503/diamonds> .Chúng tôi kiểm tra lại tên của các thuộc tính (cột) có trong tập dữ liệu:

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, màn hình

Mô tả được tạo tự động

Tiếp theo chúng tôi xem qua các thông tin chung của các thuộc tính trong bộ dữ liệu:



Có thể thấy bộ dữ liệu khá sạch khi không có trường dữ liệu ở bất kỳ sample dữ liệu nào bị thiếu (null value). Hầu hết các cột của dữ liệu ở dạng số (numerical data) và chỉ có 3 cột là dữ liệu phân loại (categorical data).

Các thuật toán học máy đều làm việc với số, nên chúng tôi quy đổi dữ liệu dạng phân loại (categorical data) về dạng số để triển khai thử nghiệm bộ dữ liệu với các thuật toán học máy.

Có một cách thường được sử dụng để giải quyết vấn đề này là sử dụng one-hot vector. Cách này tạo ra một thuộc tính nhị phân dựa trên một loại (category). Ví dụ: một thuộc tính được set là 1 khi phân loại đó là “good” (và là 0 trong trường hợp khác “good”), một thuộc tính khác được set là 1 khi cũng phân loại đó là “ideal” (và là 0 trong trường hợp khác “ideal”). Vậy nên, chúng tôi sử dụng dummy variables của thư viện pandas để chuyển đổi từ dữ liệu phân loại sang dữ liệu dưới dạng là các vecto nhị phân trong quá trình tiền xử lý trước khi đưa vào mô hình.

Tuy nhiên cách xử lý này vẫn còn một số nhược điểm mà chúng tôi sẽ nêu ra và thử nghiệm phương pháp thay thế ở phần mạng neural network trong phần sau của bài báo cáo.