Phân tích dữ liệu khám phá (EDA) là gì?

Phân tích dữ liệu khám phá (EDA) là một kỹ thuật phân tích dữ liệu mà chúng ta hiểu dữ liệu một cách chính xác. Điều quan trọng là, nó có nghĩa là hiểu những gì trong dữ liệu mà chúng ta đang làm việc. Trong bài viết này, tôi sẽ hướng dẫn bạn qua phân tích dữ liệu khám phá là gì và những bước và kỹ thuật EDA trong quá trình khoa học dữ liệu.

## Exploratory Data Analysis là gì?

Phân tích dữ liệu khám phá là bước quan trọng nhất trong bất kỳ nhiệm vụ khoa học dữ liệu nào. Các mục tiêu chính của EDA là:

* Phân tích phân phối dữ liệu
* Phát hiện các ngoại lệ và bất thường
* Lựa chọn các đặc điểm quan trọng nhất
* Loại bỏ các cột không cần thiết
* Xóa bỏ và điền vào các giá trị thiếu
* Khám phá những động cơ ẩn

Trong quá trình EDA, chúng ta cũng thực hiện lựa chọn đặc trưng và hiểu dữ liệu chủ yếu thông qua việc trực quan hóa nó, hiểu mỗi đặc điểm và phân tích mối quan hệ giữa các đặc điểm cũng là một phần quan trọng của phân tích dữ liệu khám phá.

Đây là bước đầu tiên sau khi thu thập dữ liệu và sau khi EDA, chúng ta tiến hành kỹ thuật chế biến đặc trưng và lựa chọn mô hình. Vì vậy, EDA giúp rất nhiều trong việc lựa chọn các đặc trưng tốt nhất cho mô hình và lựa chọn mô hình tốt nhất để dự đoán các nhãn.

## Các Kỹ Thuật Phân Tích Dữ Liệu Khám Phá

Hãy hiểu các kỹ thuật phân tích dữ liệu khám phá bằng cách xem xét các kỹ thuật nào nên sử dụng tùy thuộc vào loại dữ liệu chúng ta đang làm việc:

| **Loại Dữ Liệu** | **Các Kỹ Thuật EDA Nên Sử Dụng** |
| --- | --- |
| Dạng Phân Loại | Thống Kê Mô Tả |
| Univariate | Biểu Đồ Đường và Histogram |
| Bivariate | Biểu Đồ Phân Tán |
| Mảng | Bản Đồ Nhiệt |
| Đa Biến | Đám Mây Điểm 3D hoặc 2D với một biến thứ ba được biểu diễn bằng màu sắc, hình dạng hoặc kích thước khác nhau. |
| Nhiều Nhóm | Biểu Đồ Hộp |

Bây giờ hãy xem xét các kỹ thuật EDA hữu ích nhất mà chúng ta nên sử dụng tùy thuộc vào mục tiêu:

| **Mục Tiêu** | **Các Kỹ Thuật EDA Nên Sử Dụng** |
| --- | --- |
| Hiểu được phân phối của các đặc điểm. | Histogram |
| Phát hiện ngoại lệ | Histogram, biểu đồ phân tán, biểu đồ hộp |
| Hiểu mối quan hệ giữa hai biến | Biểu đồ phân tán 2D và Tương quan |
| Trực quan hóa mối quan hệ giữa hai biến đầu vào và một biến đầu vào | Bản đồ nhiệt |
| Trực quan hóa dữ liệu có số chiều cao | T-SNE hoặc PCA + Biểu đồ phân tán 2D / 3D |

## Cacs bước của EDA

Đây không phải là quy tắc chung nhưng thường được ưa chuộng khi thực hiện EDA trên bất kỳ loại bộ dữ liệu nào:

* 1. Nhập dữ liệu
* 2. Hiểu phân phối dữ liệu và giá trị thiếu
* 3. Hiểu mỗi đặc điểm
* 4. Thống kê mô tả
* 5. Hiểu tương quan giữa các đặc điểm
* 6. Phát hiện ngoại lệ

## Kết Luận

EDA là bước đầu tiên và quan trọng nhất trong bất kỳ nhiệm vụ khoa học dữ liệu nào. Bạn cần có một hiểu biết tốt về thống kê và các kỹ thuật trực quan hóa để khám phá và hiểu rõ dữ liệu của mình. Hy vọng bạn thích bài viết này về EDA trong khoa học dữ liệu.