Giới thiệu chung về khoa học dữ liệu

Trước khi đi sâu vào nó, hãy xác định khoa học dữ liệu là gì: nó là sự kết hợp của nhiều ngành, bao gồm kinh doanh, thống kê và lập trình, từ đó rút ra những hiểu biết có ý nghĩa từ dữ liệu bằng cách chạy các thí nghiệm có kiểm soát tương tự như nghiên cứu khoa học.

Mục tiêu của bất kỳ dự án khoa học dữ liệu nào là lấy được kiến thức có giá trị cho doanh nghiệp từ dữ liệu để đưa ra quyết định tốt hơn. Trách nhiệm của các nhà khoa học dữ liệu là xác định các mục tiêu cần đạt được cho một dự án. Điều này đòi hỏi kiến thức và chuyên môn kinh doanh. Mình sẽ cố gắng để các bạn tiếp cận với một số ví dụ về các nhiệm vụ khoa học dữ liệu từ các bộ dữ liệu trong thế giới thực.

Thống kê là một lĩnh vực toán học được sử dụng để phân tích và tìm kiếm các mẫu từ dữ liệu. Rất nhiều kỹ thuật mới nhất và tiên tiến nhất vẫn dựa vào các phương pháp thống kê cốt lõi. Mình sẽ trình bày cho bạn các kỹ thuật cơ bản cần có để hiểu các khái niệm sẽ đề cập.

Với sự gia tăng theo cấp số nhân trong việc tạo dữ liệu, cần nhiều sức mạnh tính toán hơn để xử lý nó một cách hiệu quả. Đây là lý do tại sao lập trình là một kỹ năng cần thiết cho các nhà khoa học dữ liệu. Bạn có thể tự hỏi tại sao đa số chúng ta sẽ chọn Python. Đó là bởi vì Python là một trong những ngôn ngữ lập trình phổ biến nhất cho khoa học dữ liệu. Thật dễ dàng để tìm hiểu cách viết mã bằng Python nhờ cú pháp đơn giản và dễ đọc của nó. Nó cũng có một số lượng đáng kinh ngạc các gói có sẵn (package) cho bất kỳ ai miễn phí, chẳng hạn như Pandas, scikit-learn, TensorFlow và PyTorch. Cộng đồng của nó đang mở rộng với một tốc độ đáng kinh ngạc, thêm ngày càng nhiều chức năng mới và cải thiện hiệu suất và độ tin cậy của nó. Không có gì lạ khi các công ty như Facebook, Airbnb và Google đang sử dụng nó như một trong những ngăn xếp chính của họ. Nếu bạn có một số kinh nghiệm với Python hoặc các ngôn ngữ lập trình khác, thì đây sẽ là một lợi thế, tất cả các khái niệm sẽ được giải thích đầy đủ, vì vậy đừng lo lắng nếu bạn chưa quen với lập trình.