**Tổng hợp kiến thức CSA - Advanced( computer scientist )**

1. Dữ liệu

* Dữ liệu là những số liệu và thống kê được tổng hợp để phân tích và đối chiếu
* Quá trình sử lý và phân tích dữ liệu thô để phát hiện xu hướng và đưa ra nhận định được gọi là phân tích dữ liệu.

1. Dữ liệu định lượng và định tính

* Dữ liệu định lượng là dữ liệu dạng số, có thể đếm và đo lường được
* Dữ liệu định tính là dữ liệu phản ánh tính chất, đặc điểm mô tả mà không thể diễn tả bằng số

1. Series và DataFrame

* Series là danh sách một chiều chứa dữ liệu, có chỉ số phần tử là một danh sách khác
* DataFrame là cấu trúc dữ liệu hai chiều, tổ chức theo dạng bảng – gồm hàng và cột

1. Trực quan hóa bằng biểu đồ

* Biểu đồ đường: Kết nối các điểm dữ liệu thành đường, biểu diễn xu hướng của một chuỗi dữ liệu.
* Biểu đồ cột: biểu diễn thông tin riêng lẻ rõ ràng, giúp người dùng tập chung vào từng giá trị riêng biệt
* Biểu đồ phân tán: là một dạng biểu đồ phân phối sử dụng các điểm để biểu diễn mối quan hệ mạnh yếu giữa các biến

1. Câu chuyện dữ liệu

* Câu chuyện dữ liệu có nhiệm vụ giải thích dữ liệu và thuyết phục khán giả về tầm quan trọng của dữ liệu
* Các bước xây dựng câu chuyện:
  + B1: Xác định khán giả
  + B2: Xác định dữ liệu
  + B3: Phác thảo cốt truyện
  + B4: Tạo bản thiết kế
  + B5: Xem lại và chỉnh sửa
  + B6: Chia sẻ câu chuyện

1. Mô hình tuyến tính

* Hồi quy tuyến tính là phương pháp mô tả mối quan hệ tuyến tính giữa hai biến.
* Đường hồi quy tuyến tính là đường thẳng vẽ khớp nhất với các điểm dữ liệu
* Tương quan chỉ độ mạnh yếu của mối quan hệ giữa các giá trị

1. Huấn luyện và đánh giá

* Huấn luyện (Training) là quá trình cho mô hình học từ một bộ dữ liệu
* Đánh giá (Testing) là quá trình sử dụng mô hình để dự đoán trên bộ dữ liệu chưa biết. Sau đó so sánh kết quả dự đoán với kết quả thực tế.
* Tỉ lệ chia tập train : test thường là 7 : 3
* Chỉ số Mean Squared Error (MSE): thể hiện mức độ lệch giữa dự đoán với thực tế. MSE càng thấp mô hình dự đoán càng chính xác.
* Chỉ số R^2: đo khoảng cách giữa các giá trị thực tế đến đường hồi quy. R^2 càng cao mô hình dự đoán càng chính xác

1. Mô hình đa biến

* Mô hình tuyến tính đa biến mô tả mối quan hệ tuyến tính giữa nhiều biến.