

REPORT

EXPLORATORY DATA ANALYST



GIỚI THIỆU : EDA LÀ GÌ?

○ ○ ○

EDA (VIẾT TẮT CỦA **EXPLORATORY DATA ANALYST**) LÀ MỘT PHƯƠNG PHÁP KHÁM PHÁ DỮ LIỆU, TÌM RA CÁC XU HƯỚNG, MẪU THỬ HOẶC KIỂM TRA CÁC GIẢ ĐỊNH TRONG DỮ LIỆU NHẰM MỤC ĐÍCH HIỂU RÕ VỀ CẤU TRÚC VÀ TÍNH CHẤT CỦA DỮ LIỆU. KHI ÁP DỤNG CÁC THUẬT TOÁN HỌC MÁY HOẶC XÂY DỰNG CÁC MÔ HÌNH DỰ ĐOÁN, EDA GÓP PHẦN QUAN TRỌNG TRONG QUÁ TRÌNH XỬ LÝ DỮ LIỆU, GIÚP GIẢI QUYẾT CÁC ĐIỀU KIỆN NGOẠI LỆ, GIÁ TRỊ THIẾU VÀ NHỮNG VẤN ĐỀ ẢNH HƯỞNG ĐẾN KẾT QUẢ CUỐI CÙNG.



MỤC ĐÍCH SỬ DỤNG?

- **TÌM HIỂU VỀ CẤU TRÚC DỮ LIỆU:** EDA LÀ PHƯƠNG PHÁP GIÚP XÁC ĐỊNH CẤU TRÚC DỮ LIỆU. KHI XÁC ĐỊNH ĐƯỢC CẤU TRÚC DỮ LIỆU, CÁC NHÀ PHÂN TÍCH DỮ LIỆU CÓ THỂ HIỂU ĐƯỢC MỐI QUAN HỆ GIỮA CÁC DỮ LIỆU TRONG TẬP.
- **ĐIỀU CHỈNH VÀ THAY ĐỔI:** EDA GIÚP GIẢI QUYẾT CÁC TRƯỜNG HỢP THIẾU GIÁ TRỊ, DỮ LIỆU LỖI, CÁC NGOẠI LỆ TRONG DỮ LIỆU. ĐIỀU NÀY GIÚP CÁC NHÀ PHÂN TÍCH DỮ LIỆU ĐIỀU CHỈNH CÁC PHƯƠNG ÁN KHẮC PHỤC KỊP THỜI, TRÁNH NHỮNG ẢNH HƯỞNG NGHIÊM TRỌNG ĐẾN DỰ ÁN.



- XÁC ĐỊNH MỐI TƯƠNG QUAN GIỮA CÁC BIẾN:** CÁC BIẾN ĐỀU CHỨA CÁC GIÁ TRỊ RIÊNG, EDA CÓ KHẢ NĂNG PHÁT HIỆN CÁC LIÊN HỆ TIỀM ẨN VÀ SỰ ẢNH HƯỞNG GIỮA CÁC BIẾN VỚI NHAU, TẠO SỰ LIÊN KẾT GIỮA CÁC THÔNG TIN DỮ LIỆU, XÂY DỰNG MỘT QUY TRÌNH PHÂN TÍCH TỔNG THỂ, RÕ RÀNG.
- XÂY DỰNG CƠ SỞ DỮ LIỆU QUAN HỆ:** CẤU TRÚC HÓA DỮ LIỆU THEO SƠ ĐỒ, TIẾT KIỆM THỜI GIAN XỬ LÝ NHỮNG THÔNG TIN THỪA, HẠN CHẾ SỰ SAO SÓT CỦA KẾT QUẢ PHÂN TÍCH.
- CHUẨN BỊ CHO BƯỚC PHÂN TÍCH TIẾP THEO:** ÁP DỤNG EDA GIÚP LOẠI BỎ CÁC DỮ LIỆU KHÔNG CẦN THIẾT, DỮ LIỆU THIẾU GIÁ TRỊ VÀ CHUẨN HÓA DỮ LIỆU. ĐÂY LÀ YẾU TỐ NỀN TẢNG ĐỂ CHUẨN BỊ CHO CÁC BƯỚC PHÂN TÍCH BẰNG THUẬT TOÁN HỌC MÁY.

MỤC ĐÍCH SỬ DỤNG?

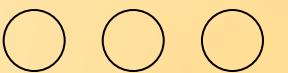




KỸ THUẬT



PHÂN TÍCH ĐƠN BIẾN



PHÂN TÍCH ĐƠN BIẾN ĐƯỢC THỰC HIỆN VỚI MỤC ĐÍCH LÀ HIỂU ĐƯỢC SỰ PHÂN BỐ CỦA CÁC GIÁ TRỊ CHO MỘT BIẾN DUY NHẤT. DỮ LIỆU ĐƠN BIẾN KHÔNG THEO LOẠI DỮ LIỆU CỤ THỂ MÀ ĐƯỢC PHÂN THEO MỤC ĐÍCH SỬ DỤNG HOẶC BẢN CHẤT RIÊNG. ĐỂ PHÂN TÍCH MỘT TẬP DỮ LIỆU, CÁC LOẠI KỸ THUẬT PHÂN TÍCH ĐƠN BIẾN SẼ ĐƯỢC SỬ DỤNG TÙY THUỘC VÀO CÁC LOẠI BIẾN ĐỀ CẬP. MỘT SỐ DẠNG BIỂU ĐỒ ĐƯỢC SỬ DỤNG NHIỀU TRONG PHÂN TÍCH ĐƠN BIẾN :

- **HISTOGRAMS** (BIỂU ĐỒ PHÂN PHỐI)
- **BOXPLOT** (BIỂU ĐỒ HỘP)
- **BAR CHART** (BIỂU ĐỒ CỘT)
- **PIE CHART** (BIỂU ĐỒ TRÒN)



PHÂN TÍCH HAI BIỂN

○ ○ ○

PHÂN TÍCH HAI BIỂN LÀ PHƯƠNG PHÁP KIỂM TRA SỰ LIÊN QUAN GIỮA HAI DỮ LIỆU KHÁC NHAU. ĐÂY LÀ MỘT KỸ THUẬT PHÂN TÍCH GIÚP XÁC ĐỊNH CÁCH KẾT NỐI GIỮA HAI BIỂN VÀ TÌM RA XU HƯỚNG TRONG DỮ LIỆU. CÁC DẠNG BIỂU ĐỒ PHỔ BIỂN ĐƯỢC SỬ DỤNG CHO PHÂN TÍCH HAI BIỂN NHƯ:

- **SCATTER PLOT** (BIỂU ĐỒ PHÂN TÁN) : BIỂU ĐỒ PHÂN TÁN CHO BIẾT HAI BIỂN CÓ LIÊN QUAN NHƯ THẾ NÀO. THỂ HIỆN CÁC GIÁ TRỊ CỦA MỘT BIỂN TRÊN TRỤC X VÀ CÁC GIÁ TRỊ KHÁC CỦA BIỂN TRÊN TRỤC Y.
- **CORRELATION** (BIỂU ĐỒ TƯƠNG QUAN) : HỆ SỐ TƯƠNG QUAN LÀ PHÉP ĐO THỂ HIỆN MỨC ĐỘ MẠNH VÀ ĐỊNH HƯỚNG CỦA HAI BIỂN ĐƯỢC LIÊN KẾT.
- **REGRESSION** (BIỂU ĐỒ PHÂN TÍCH HỒI QUY) : TRONG BIỂU ĐỒ HỒI QUY, TRỤC X ĐẠI DIỆN CHO BIỂN ĐỘC LẬP VÀ TRỤC Y ĐẠI DIỆN CHO BIỂN PHỤ THUỘC. KHI CÁC ĐIỂM ĐƯỢC THỂ HIỆN TRÊN BIỂU ĐỒ, MỘT ĐƯỜNG HỒI QUY SẼ ĐƯỢC VẼ ĐỂ ƯỚC LƯỢNG MỐI QUAN HỆ TUYẾN TÍNH GIỮA HAI BIỂN.



PHÂN TÍCH ĐA BIẾN



PHÂN TÍCH ĐA BIẾN KỸ THUẬT PHÂN TÍCH Ở CẤP ĐỘ PHỨC TẠP HƠN, ĐƯỢC SỬ DỤNG KHI CÓ NHIỀU HƠN HAI BIẾN TRONG TẬP DỮ LIỆU. PHÂN TÍCH ĐA BIẾN GIÚP GIẢM THIỂU VÀ ĐƠN GIẢN HÓA DỮ LIỆU MÀ KHÔNG LÀM MẤT BẤT KỲ CHI TIẾT QUAN TRỌNG NÀO TRONG TẬP DỮ LIỆU.

- **MULTIVARIATE DISTRIBUTION PLOT:** BIỂU ĐỒ NÀY CHO PHÉP XEM KẾT QUẢ PHÂN PHỐI CỦA NHIỀU BIẾN ĐỒNG THỜI, ĐƯỢC HIỂN THỊ DƯỚI DẠNG 2D HOẶC 3D VỀ MẬT ĐỘ PHÂN PHỐI VÀ SỰ MỐI LIÊN HỆ GIỮA CÁC BIẾN.
- **SCATTERPLOT MATRIX:** BIỂU ĐỒ THỂ HIỆN ĐƯỢC SỰ PHÂN PHỐI VÀ MỐI QUAN HỆ GIỮA CÁC BIẾN DO CÁC CẶP TƯƠNG QUAN GIỮA CÁC BIẾN ĐÃ ĐƯỢC HIỂN THỊ ĐẦY ĐỦ.
- **BOXPLOT MATRIX:** BIỂU ĐỒ THỂ HIỆN SỰ PHÂN BỐ ĐẶC TRƯNG CỦA CÁC BIẾN ĐỒNG THỜI ĐỂ SO SÁNH VÀ TÌM RA ĐIỂM KHÁC BIỆT GIỮA CÁC DỮ LIỆU.

QUY TRÌNH

- **BƯỚC 1 - THU THẬP DỮ LIỆU**
- **BƯỚC 2 - KIỂM TRA DỮ LIỆU**
- **BƯỚC 3 - XỬ LÝ DỮ LIỆU**
- **BƯỚC 4 - TRỰC QUAN DỮ LIỆU**
- **BƯỚC 5 - ĐÚC KẾT**
- **BƯỚC 6 - BÁO CÁO KẾT QUẢ**



THANK

YOU

