ĐỀ ÁN DỰ ÁN ĐÁNH GIÁ NĂNG LỰC INTERN DATA ANALYST p2

Dự án: PHÂN TÍCH HIỆU QUẢ BÁN HÀNG VÀ DỰ ĐOÁN DOANH THU (AdventureWorks2022)

1. Mục tiêu dự án

Dự án nhằm đánh giá toàn diện năng lực của Intern Data Analyst:

- Truy vấn, tổng hợp dữ liệu bán hàng bằng **SQL** từ cơ sở dữ liệu AdventureWorks2022.
- Phân tích dữ liệu, tìm insight và xây dựng mô hình dự đoán doanh thu bằng **Python**.
- Trực quan hóa dữ liệu động, báo cáo quản trị bằng **Power BI**.
- Viết báo cáo phân tích và đề xuất cải thiện doanh thu bằng **Word/PDF**.

---

2. Mô tả dữ liệu AdventureWorks2022 (một số bảng chính)

Nhóm dữ liệu đơn hàng

- **Sales.SalesOrderHeader**: Thông tin chung của đơn hàng
  - SalesOrderID, OrderDate, ShipDate, DueDate, CustomerID, SubTotal, TaxAmt, Freight, TotalDue, TerritoryID
- **Sales.SalesOrderDetail**: Chi tiết sản phẩm trong mỗi đơn hàng
  - SalesOrderID, ProductID, OrderQty, UnitPrice, UnitPriceDiscount, LineTotal

Nhóm dữ liệu khách hàng

- **Sales.Customer**: Loại khách hàng (Individual hoặc Store)
  - CustomerID, PersonID, StoreID, TerritoryID
- **Person.Person**: Thông tin cá nhân khách hàng (FirstName, LastName, Email, PhoneNumber)

Nhóm dữ liệu sản phẩm

- **Production.Product**: Thông tin sản phẩm
  - ProductID, Name, ProductNumber, StandardCost, ListPrice, Category

Nhóm dữ liệu lãnh thổ bán hàng

- **Sales.SalesTerritory**: Thông tin khu vực bán hàng
  - TerritoryID, Name, CountryRegionCode

---

3. Yêu cầu dự án

3.1. Phần 1: Khai thác dữ liệu (**SQL**)

1. Kết nối tới cơ sở dữ liệu AdventureWorks2022 (SQL Server).
2. Viết câu lệnh SQL để:
   - Lấy danh sách **doanh thu theo năm và khu vực bán hàng**.
   - Tìm **Top 10 sản phẩm bán chạy nhất theo số lượng và doanh thu**.
   - Tính **tỷ lệ hoàn thành giao hàng đúng hạn** theo Territory.
   - Tính **giá trị trung bình đơn hàng** theo từng loại khách hàng (Individual vs Store).
3. Làm sạch dữ liệu ở mức cơ bản (xử lý NULL, loại bỏ bản ghi trùng).
4. Lưu lại các câu lệnh SQL (*.sql).

3.2. Phần 2: Tiền xử lý và phân tích khám phá dữ liệu (**Python**)

1. Kết nối cơ sở dữ liệu hoặc import dữ liệu đã xuất từ SQL.

2. Thực hiện EDA:
    - Thống kê doanh thu, số lượng đơn hàng theo thời gian, sản phẩm, khu vực.
    - Phân tích xu hướng doanh thu theo mùa vụ (tháng, quý).
    - Xác định các yếu tố ảnh hưởng lớn nhất đến doanh thu (sản phẩm, khu vực, loại khách hàng).
3. Xuất báo cáo EDA sang file PDF/Word kèm biểu đồ minh họa.

3.3. Phần 3: Xây dựng mô hình Machine Learning (**Python**)

1. Mục tiêu: Dự đoán **doanh thu đơn hàng (TotalDue)** dựa trên các biến đầu vào (sản phẩm, số lượng, loại khách hàng, khu vực).
2. Thực hiện:
    - Làm sạch dữ liệu và chọn đặc trưng phù hợp.
    - Chia dữ liệu train/test.
    - Xây dựng ít nhất 3 mô hình ML (Linear Regression, Random Forest, Gradient Boosting).
    - Đánh giá bằng RMSE, MAE, R^2.
    - Chọn mô hình tốt nhất và giải thích lý do.

3.4. Phần 4: Trực quan hóa dữ liệu và kết quả (**Power BI**)

1. Tạo dashboard động hiển thị:
    - Doanh thu theo năm/quý/tháng và khu vực bán hàng.
    - Top 10 sản phẩm bán chạy.
    - Tỷ lệ giao hàng đúng hạn theo Territory.
    - Kết quả dự đoán doanh thu từ mô hình ML.
2. Dashboard phải có bộ lọc tương tác (filter theo khu vực, thời gian, loại khách hàng).
3. **Lưu ý dữ liệu phải được kết nối từ sql server**

3.5. Phần 5: Báo cáo cuối cùng (**Word/PDF**)

Báo cáo cần gồm:

1. Mục tiêu và phương pháp.
2. Tóm tắt dữ liệu đã dùng và các bước xử lý.
3. Kết quả phân tích EDA.
4. Kết quả mô hình ML.
5. Dashboard và insight rút ra.
6. Đề xuất cải thiện doanh thu cho công ty.