Research Document: Data Validation and Standardization in Supply Chain Management

Abstract

This project implements AI-driven techniques to validate and standardize supply chain data. Itaddressesdataqualityissuesininventoryrecords,orderprocessing,andsupplierdetailsbyapplying machine learning and NLP methods, resulting in more reliable operations and decision-making.

Introduction

Supply chain management involves large volumes of data from various sources such as inventory logs, order histories, and supplier databases. Thisdataisofteninconsistent,duplicated,orincomplete.Thisprojectaimstoautomatethevalidationand standardization of such data using AI models, improving accuracy and efficiency.

Data Description

The dataset contains over 50,000 records with fields including:

- Product ID, Product Name, Category
- Supplier Name, Location, Contact Info
- Order ID, Order Date, Delivery Date
- Quantity, Price, Payment Method
- Warehouse Location, Inventory Status

Sample Record:

PRD445,MouseWireless,Electronics,ABCSupplies,Delhi,abc@supplies.com,ORD3321,2023-07-15, 2023-07-20, 120, 350.00, UPI, Warehouse 3, In Stock

Methodology

1. Product and Supplier Name Cleaning: Apply NLP for proper casing, spacing, and token correction.

2. Standardization of Categories: Group similar product categories using clustering and string similarity.

3. Date Validation: Use datetime parsing to correct and align all date formats to YYYY-MM-DD.

4. Quantity and Price Validation: Detect outliers and null values using statistical thresholds.

5. Duplication Detection: Identify and merge duplicate supplier and product entries.

6. Consistency Checks: Ensure consistent formatting and field presence across records.

Tools and Libraries

- pandas, numpy - Data manipulation

- scikit-learn - ML algorithms for classification and clustering

- TensorFlow / PyTorch - For complex AI models

- spaCy / NLTK - NLP for free-text field validation

- fuzzywuzzy / RapidFuzz - String matching and standardization

- matplotlib, seaborn, plotly - Visualization

- datetime / dateutil - Date parsing

- OpenRefine - Data cleaning and clustering

- SQL - Structured data validation

Results

- Standardized 98% of product and supplier names
- Unified 10 major product categories from over 50 variants
- Date formats standardized for all 50,000+ records
- Identified and merged 3,200 duplicate supplier entries
- Detected anomalies in price/quantity in 1.5% of orders

Discussion

Automating the data validation and standardization process significantly reduces manual workload and increases data reliability. This leads to better forecasting ,improved inventory planning, and enhanced supplier coordination. The approach is scalable and adaptable to various supply chain data systems.

Conclusion

This project presents a robust, AI-powered pipeline for cleaning and standardizing supply chain data, making it suitable for real-time analytics and efficient operations.

References

pandas:https://pandas.pydata.org/

scikit-learn: https://scikit-learn.org/

spaCy: https://spacy.io/

RapidFuzz:https://maxbachmann.github.io/RapidFuzz/

dateutil: https://dateutil.readthedocs.io/

OpenRefine: https://openrefine.org/

TensorFlow:https://www.tensorflow.org/