Project Documentation: E-commerce Product Analysis using Machine Learning

# 1. Introduction

E-commerce is one of the fastest-growing domains globally, driven by data-rich platforms and ever-changing consumer behavior. This project aims to harness the power of data science to analyze product data scraped from an e-commerce website, enabling smarter decisions regarding product offerings, segmentation, and classification.

# 2. Aim

The goal of this project is to implement a complete end-to-end data science pipeline—from web scraping to machine learning—for analyzing e-commerce product data. The insights aim to help businesses enhance their marketing strategies, pricing models, and customer targeting efforts.

# 3. Business Problem / Problem Statement

In a competitive online marketplace, understanding product performance, customer preferences, and market trends is critical. E-commerce platforms often struggle to segment products meaningfully or predict their success. This project seeks to:

- Cluster products based on shared characteristics (ratings, reviews, pricing).

- Predict product categories based on key features.

- Optimize product recommendations and business strategy through data insights.

# 4. Project Workflow

1. Web Scraping using Python

2. Data Cleaning and Preprocessing

3. Feature Engineering

4. Exploratory Data Analysis (EDA)

5. Unsupervised Learning (Product Clustering)

6. Supervised Learning (Category Prediction)

7. Hyperparameter Tuning

8. Model Evaluation and Documentation

# 5. Data Collection

Web scraping was performed on an e-commerce website using BeautifulSoup or Scrapy to collect:

- Product Name

- Price

- Category

- Ratings

- Number of Review

The data was saved in CSV format for further analysis.

# 6. Data Cleaning

- Removed missing values and duplicates

- Standardized price formats and text cases

- Converted ratings and review counts to numeric values

- Filtered irrelevant or inconsistent data entries

# 7. Feature Engineering

- Extracted numerical features from text-based data

- Converted categorical data using label encoding

- Derived product metrics such as price buckets and review quality

- Created new features for clustering and modeling

# 8. Exploratory Data Analysis (EDA)

Performed visual analysis to understand:

- Distribution of prices and ratings

- Relationship between reviews and ratings

- Category-wise product trends

- Correlation between features

Tools used: matplotlib, seaborn, and pandas profiling

# 9. Unsupervised Learning (Product Clustering)

Applied K-Means Clustering to group similar products based on:

- Price

- Ratings

- Review counts

- Used the Elbow Method to determine the optimal number of clusters

- Added cluster labels to the dataset for business segmentation

# 10. Supervised Learning (Category Prediction)

Trained multiple classification models to predict product categories:

- Logistic Regression

- Support Vector Machine (SVM)

- K-Nearest Neighbors (KNN)

- Random Forest

- XGBoost

Used accuracy, F1-score, and confusion matrix to evaluate each model. Random Forest and XGBoost provided the best performance.

# 11. Hyperparameter Tuning

Implemented GridSearchCV and RandomizedSearchCV to fine-tune models, focusing on:

- Number of trees in forest

- Maximum depth

- Learning rate (for XGBoost)

- Kernel and C values (for SVM)

Tuned models improved overall accuracy by ~10–15%.

# 12. Insights from Analysis

- Mid-range priced products received more reviews and better ratings.

- Clustering helped identify high-performing and underperforming product segments.

- Model predictions can automate product categorization, saving manual efforts.

- Product categories and price segments showed strong correlation patterns.

# 13. Conclusion

This project demonstrates a full-cycle machine learning application in the e-commerce domain. It highlights how data collection, analysis, and modeling can generate valuable business insights. The clustering and classification models not only support decision-making but can also be integrated into product recommendation systems or dashboards for real-time analytics.