-Final project-

Title: Women's Clothing E-Commerce Project
Author: Sanaullh Shafaq - Nurgazt Dias Aslanuly
Institution:CS
Course: Data Mining
Date:5/12/2025

Link Git-Hub: https://github.com/Sanaullah-cd/Women-s-Clothing-E-Commerce-.git

1. Introduction
   Purpose: This project analyzes Women's Clothing E-Commerce Reviews to understand customer sentiment, product feedback, and recommendation trends.
   Goals: Explore data, preprocess it, build models (ML, NLP, NN), analyze results, and provide insights.

Figure 1: Data lifecycle and stages of analysis.

2. Data Description and Collection
   Source: Kaggle / CSV file 'Womens Clothing E-Commerce Reviews.csv'
   Number of records: X
   Number of variables: Y
   Data types: Numerical (Rating, Age, Feedback), Categorical (Department, Division, Class), Text (Review Text)

Figure 2: Dataset structure and variable types.

3. Data Preprocessing

- Handle missing values: Review Text filled with empty string; Recommended IND derived from Rating.
- Remove duplicates.
- Normalize numeric features.
- Encode categorical features.
- Split data into train/test sets.

Figure 3: Missing values before and after cleaning.
Figure 4: Normalized numerical features.

4. Exploratory Data Analysis (EDA)

- Statistical summary and distributions of Age, Rating, and Recommendations.
- Visualizations: Rating histogram, Age histogram, Department barplots, Recommended pie chart.

Figure 5: Correlation heatmap of numerical variables.
Figure 6: Boxplot showing feature distributions.

5. Statistical Analysis and Hypothesis Testing

- ANOVA test on Rating across top 3 Departments.
- Key finding: Significant differences observed (p-value = ...).

Figure 7: Linear regression fit and confidence interval.

6. Machine Learning Models
   6.1 Supervised Learning

- Logistic Regression (TF-IDF text features), Random Forest (numeric features).
- Metrics: Accuracy, Precision, Recall, F1-score.

Figure 8: Confusion matrix of classification results.

6.2 Unsupervised Learning

- KMeans clustering on numeric features.
- PCA for dimensionality reduction.

Figure 9: PCA 2D visualization of clusters.

7. Time Series Analysis

- Synthetic monthly review counts generated.
- Seasonal decomposition and ARIMA model applied.

- Forecast evaluation: MAE, RMSE.

Figure 10: Time series forecasting plot (Predicted vs Actual).

8. Neural Network and Deep Learning

- Simple dense NN with input, hidden, and output layers for numeric features.
- Trained for 10 epochs.
- Compared with traditional ML models.

Figure 11: Neural network architecture diagram.
Figure 12: Loss and accuracy curves during training.

9. Natural Language Processing (NLP)

- Text cleaning, tokenization.
- TF-IDF vectorization.
- Word cloud visualization.
- Sentiment distribution analysis.

Figure 13: Word cloud of most frequent terms.
Figure 14: Sentiment distribution chart.

10. Ethics and Data Security

- Bias detection: Age, Department categories.
- Fairness analysis: Recommendations.
- Privacy: Users' personal reviews.

Table 1: Potential ethical risks and mitigation strategies.

11. Results and Discussion

- Model performance summary.
- TF-IDF+LogReg, RandomForest, Neural Network accuracies.
- Insights: TF-IDF performs well on text, RandomForest and NN comparable on numeric data.

Figure 15: Model performance comparison (bar chart).

12. Conclusion and Future Work

- Learned: Data patterns, sentiment, and predictive modeling.
- Improvements: Larger dataset, full deep learning on text, better hyperparameter tuning.

13. References

- Dataset: Kaggle "Women's Clothing E-Commerce Reviews"
- Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, nltk, tensorflow, wordcloud
- Documentation: scikit-learn, TensorFlow, statsmodels