Objective→ Hypothesis:

The objective of this project is to examine the Women's Clothing E-Commerce dataset to better understand the elements that influence consumer comments and recommendations.

Hypothesis: I believe that customer age, product rating, department name, and class name will all have a substantial impact on both recommendation and positive feedback counts.

Data Description:

This dataset contains customer reviews from a Women's Clothing E-Commerce site, covering features like:  
  
Clothing ID: The identifier for the individual clothing item being assessed.  
Age: The age of the reviewer.  
Rating: The customer's rating of the product (ranging from 1 to 5).  
Department Name: The product department's category name.  
Class Name: The categorical name of the product class.  
Recommended IND: A binary variable that indicates if the consumer suggested the product (1 = recommended, 0 = not recommended).  
Positive Feedback Count: The number of other customers that rated the review positively.

Describing My Process

I started by loading the dataset and performing data preprocessing operations, such as removing irrelevant columns ('Title', 'Review Text', 'Division Name', 'Unnamed: 0') and managing missing values. Categorical variables were encoded by using LabelEncoder.  
  
I used different graphs to find insights into the distribution and interactions between variables. For example, I made a histogram to show the distribution of customer age and a bar plot to look at the relationship between product rating and total positive feedback count. In addition, I used a heatmap to look into any connections between different variables in the dataset.

For classification tasks, I trained Logistic Regression, KNeighbors, SVM, Random Forest, and Decision Tree models to predict the 'Recommended IND' column. Confusion matrices were created to assess model performance.

For regression tasks, I used Linear Regression, KNeighbors Regression, Random Forest Regression, and Decision Tree Regression models to forecast the 'Positive Feedback Count'. Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were used to evaluate model correctness.

Throughout the process, I faced the following issues:

1. Data Cleaning: The dataset contains missing values, so I had to remove rows with missing values in the 'Department Name' and 'Class Name' columns.
2. Complications: Understanding the relationships between variables and determining the most important elements for prediction proved difficult.

Observations

For classification, Logistic Regression had the best accuracy (93.33%), suggesting its usefulness in predicting whether customers would recommend a product. This suggests that age, rating, department name, and class name are valid indicators of customer preferences. KNeighbors, SVM, RandomForest, and Decision Tree models all performed well, although Logistic Regression and SVM outperformed them significantly. This implies that the linear relationship depicted by Logistic Regression may be appropriate for this classification task.

For regression, Linear Regression produced the lowest Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), showing that it performed better in predicting positive feedback counts. This implies a correlation between age, rating, department name, class name, and favorable feedback received. KNN Regression, Random Forest Regression, and Decision Tree Regression all produced good predictions, albeit with somewhat higher MSE and RMSE than Linear Regression. This suggests that, while these models capture certain nonlinear interactions, they are less effective than Linear Regression in this scenario. The relatively low MSE and RMSE of all regression models indicate that age, rating, department name, and class name are significant predictors of positive feedback count, emphasizing their importance in customer satisfaction and product evaluation.

Conclusion + Future

Ultimately, the analysis of the Women's Clothing E-Commerce dataset demonstrated that customer age, product rating, department name, and class name all have an impact on both product suggestions and positive feedback counts. The classification and regression models based on these variables showed encouraging results, with some algorithms going above others in terms of accuracy and predictive power.

For the future, I could:

1. Feature importance: Knowing which variables have the most influence on customer recommendations and positive feedback counts can help guide focused marketing efforts and product developments.
2. Model refinement: Fine-tuning model hyperparameters and experimenting with advanced tools such as ensemble methods can assist with improving forecast accuracy.
3. Integrating external data: By including additional datasets like as demographic information or customer behavior data, you can gain a better understanding of customer preferences and habits.
4. Sentiment analysis: Using natural language processing techniques to evaluate review material may capture complex customer attitudes and improve prediction models.