PREDICTING CHOCOLATE BRAND PREDICTION USING WEB SCRAPING AND MACHINE LEARNING

|  |
| --- |
| NAME: JEFRIN SOBI S |
| COURSE: DA/DS, Offline |
| BATCH: JUNE 2025 |

# INTRODUCTION:

This project focuses on predicting chocolate brand categories using machine learning techniques. By performing web scraping from e-commerce websites, a dataset containing information such as brand, flavor, type, ingredients, rating, discount, and selling price was collected. The data was then cleaned, analyzed, and used to train machine learning models to predict the chocolate brand or price category.

# AIM:

* To predict the chocolate brand or category using web-scraped product data.
* To analyze the relationship between product features such as price, rating, and discount.
* To build and evaluate machine learning models for accurate prediction.

# PROBLEM STATEMENT:

With hundreds of chocolate brands available online, customers often struggle to identify the best value products. This project aims to use machine learning to classify chocolates into various categories or predict their brands based on key attributes such as ingredients, type, price, and discount.

.

# PROJECT WORKFLOW:

 **Web Scraping:** Collected data from e-commerce platforms.

 **Data Cleaning:** Removed missing or duplicate values.

 **Feature Engineering:** Extracted and derived meaningful features such as “Discount %”, “Price Range”, and “Brand Category”.

 **Exploratory Data Analysis (EDA):** Visualized relationships between features like Rating, Price, and Discount.

 **Data Preprocessing:** Encoded categorical variables and scaled numeric features.

 **Model Training:** Trained models such as Logistic Regression, Decision Tree, Random Forest, and XGBoost.

 **Model Evaluation:** Compared model performance using accuracy and F1-score.

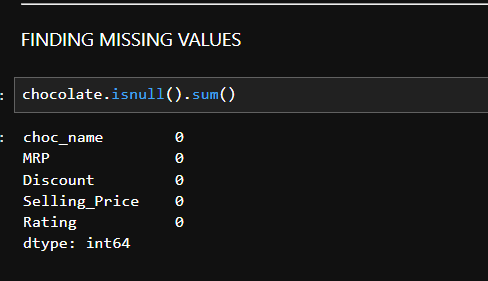
 **Deployment (Optional):** Integrated prediction functionality into a **Streamlit app**.

# DATA UNDERSTANDING:

The dataset was built using web scraping and contains information such as:  
- Brand  
- Chocolate Name **-** Flavor / Type (e.g., Milk, Dark, White, Caramel)  
- Discount (%)  
- Selling Price (₹)  
- Rating  
- MRP Price  
  
Target Variable: Chocolate Brand or Price Category

DATA CLEANING:

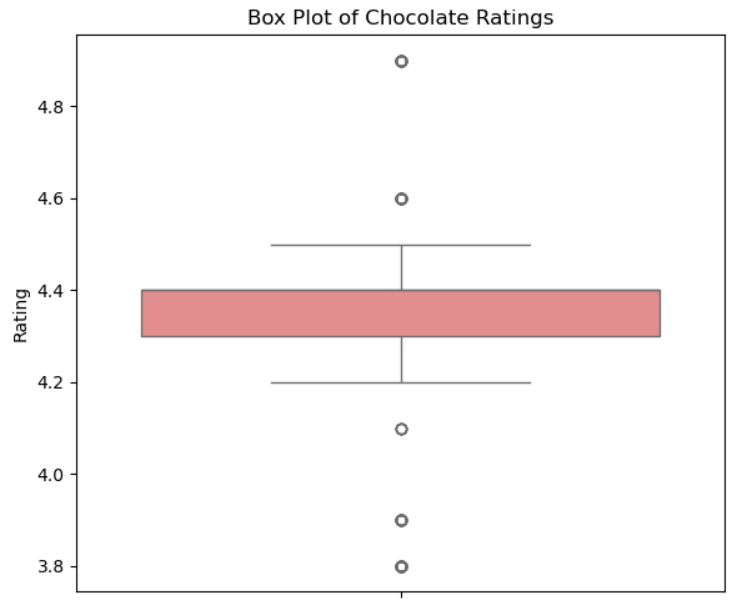
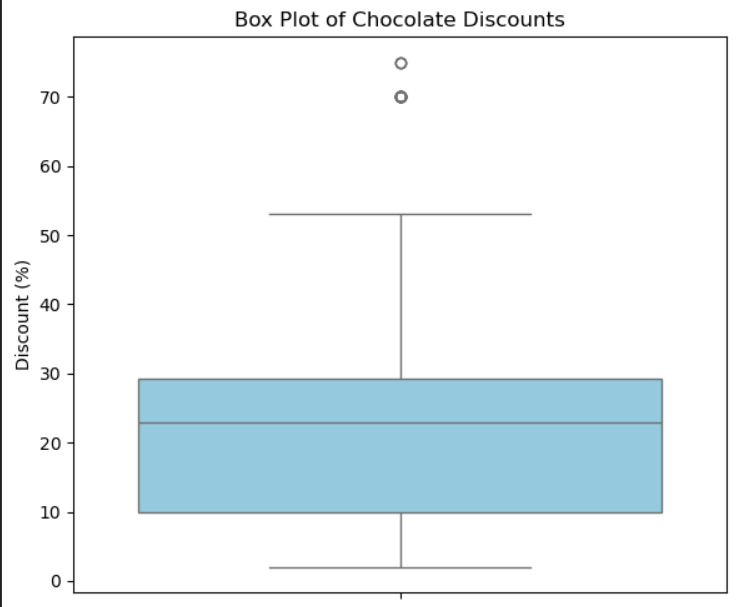
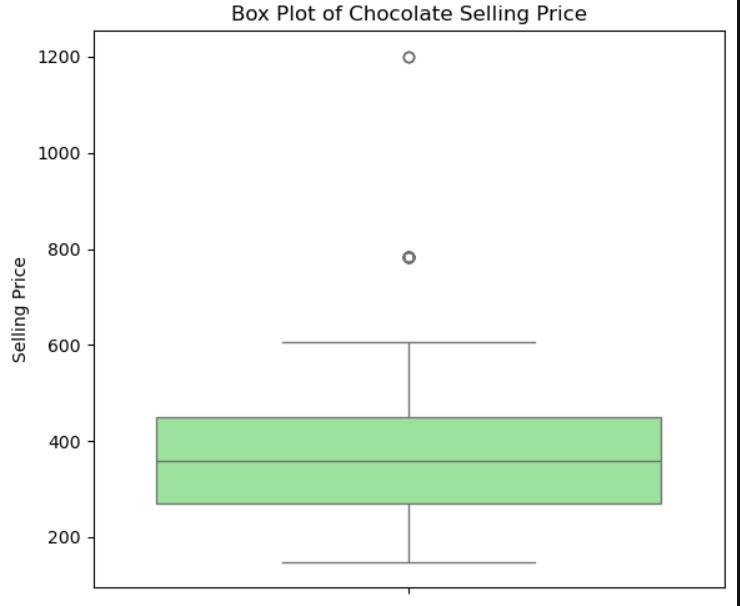
Handled missing and inconsistent values, removed duplicate entries, and standardized categorical names (e.g., brand names). Numeric columns were scaled, and textual data such as product descriptions were cleaned using regular expressions.



# 7. FEATURE ENGINEERING

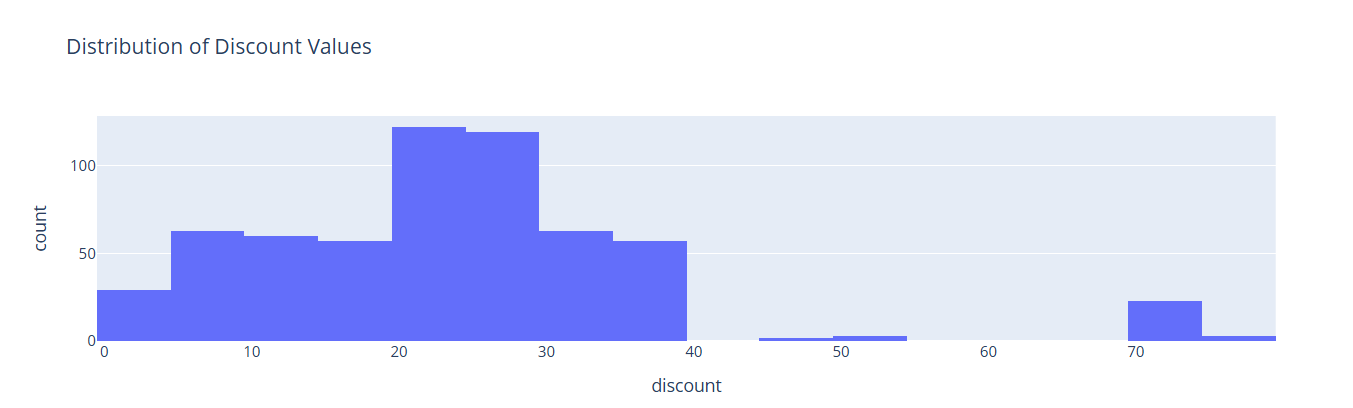
Derived new features such as:  
- Price Range (Low / Medium / High) based on selling price.  
- Discount Category (Low, Moderate, High).  
- Feature Count: Total descriptive attributes per chocolate  
These enhanced the predictive ability of the model.

# 8. OUTLIERS:



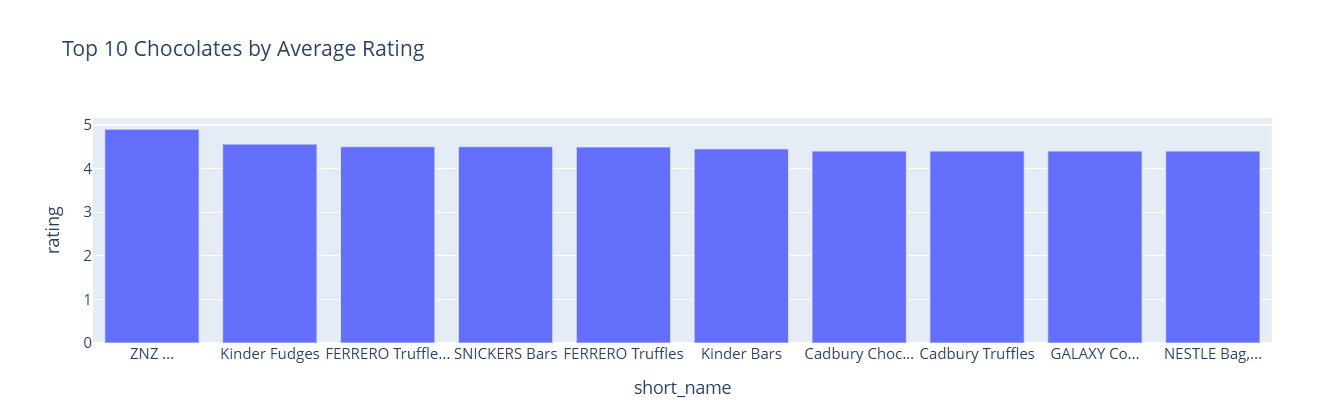
# 9.EXPLORATORY DATA ANALYSIS (EDA):

**Univariate Analysis:**  
- Distribution of discount values of chocolate



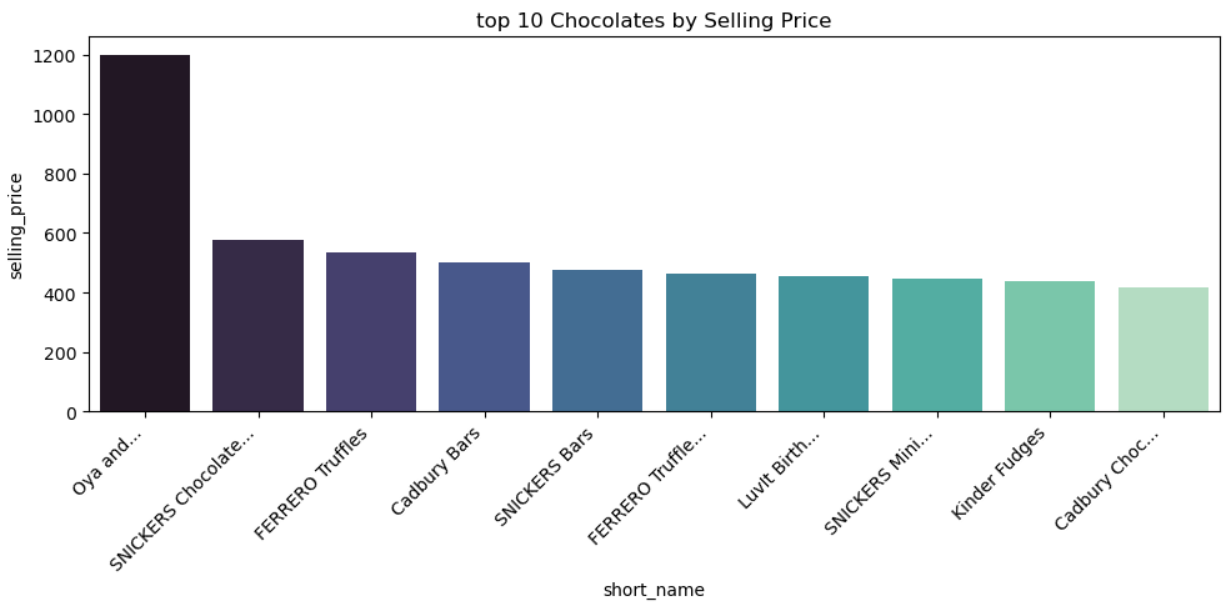
* Most chocolates have a discount between 10% and 30%, indicating this is the common promotional range.
* The distribution is right-skewed, meaning a few items have extremely high discounts compared to the majority.
* Very few chocolates have high discounts (above 50%), suggesting such offers are rare.
* A smaller number of products have low or no discounts (0–5%), likely premium or high-demand chocolate.

**Bivariate Analysis**

- Top 10 chocolates by Average Rating 

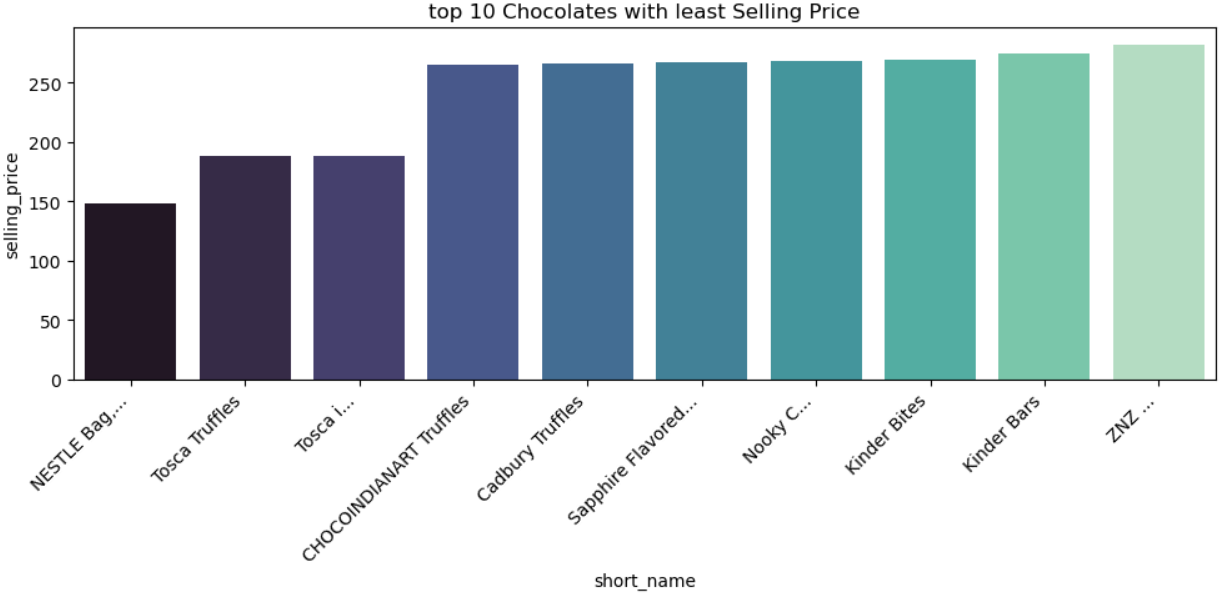
* ZNZ chocolates have the highest customer rating (4.9) — customers love this brand the most
* Kinder Fudges and FERRERO Truffles follow closely with ratings around 4.5–4.56, showing strong brand satisfaction.
* Cadbury and GALAXY products maintain consistent ratings (around 4.4), showing steady brand trust.
* The ratings difference among top brands is small (4.4 to 4.9) — this means most popular chocolates are highly rated and have similar customer satisfaction.

-Top 10 chocolates by selling price

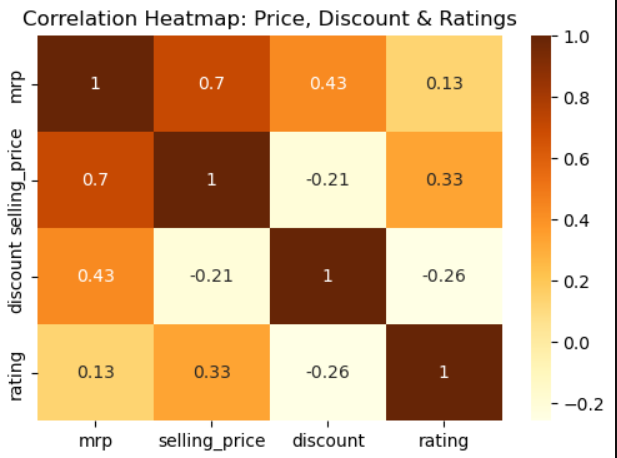


* Oya and... has the highest selling price (~₹1200) — much higher than others.
* SNICKERS and FERRERO chocolates are consistently among the top premium-priced items
* Cadbury Bars and LuvIt are in the mid-range of the top-priced chocolates.
* Prices drop sharply after the first brand — indicating a large price gap between the most expensive and others.

-Top 10 chocolates with least selling price

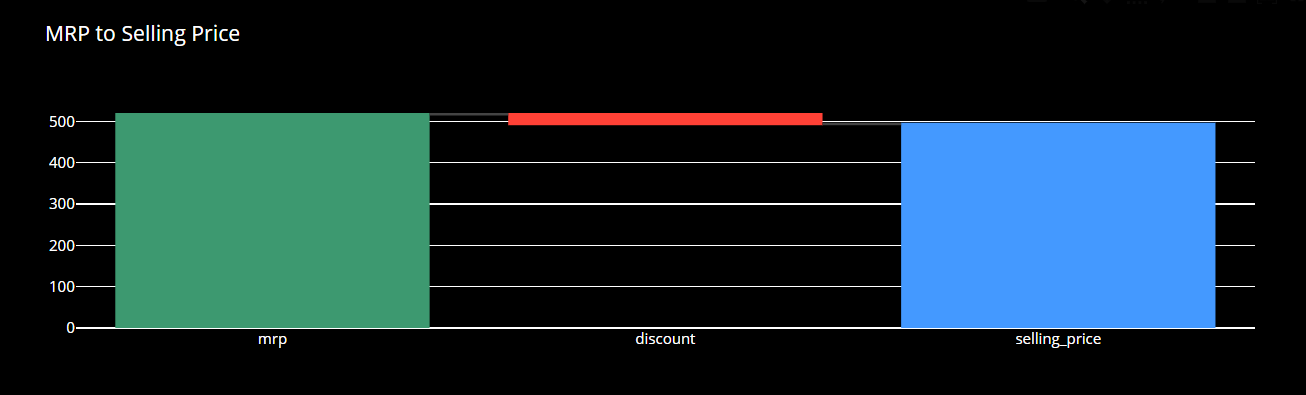


* Most chocolates are priced between ₹200–₹800 MRP, with selling prices typically lower indicating frequent discounts.
* Larger circles = higher discount. → So, bigger circles represent chocolates with big discounts (possibly seasonal or promotional offers).
* Higher MRP chocolates (₹1500–₹2500) are heavily discounted — luxury brands likely reduce price to attract buyers.
* **Multivariate Analysis:**  
  - MRP and Selling Price of chocolate have high correlation



* MRP and Selling Price have a strong positive correlation (0.7) → higher MRP → higher selling price.
* Discount has a moderate positive correlation with MRP (0.43) → expensive chocolates tend to get slightly higher discounts.
* Discount is negatively correlated with Rating (-0.26) → higher discounts might come on lower-rated chocolates.

-Waterfall chart to show mrp to selling price



## ****10. UNSUPERVISED LEARNING:****

### In this part of the project, unsupervised learning techniques were used to group chocolates into clusters based on their key features such as selling price, discount, rating, and brand type. The aim was to discover natural groupings and patterns among chocolates without using brand or price labels.

**Methods Used**

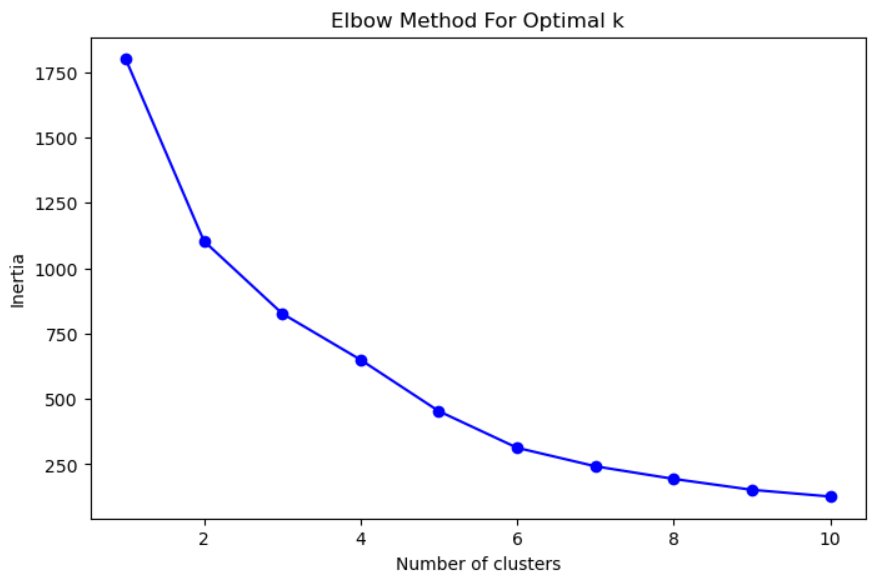
#### ****1. K-Means Clustering:****

Applied **K-Means Clustering** to group chocolates based on features like price, discount, and rating.

**Clusters formed:**

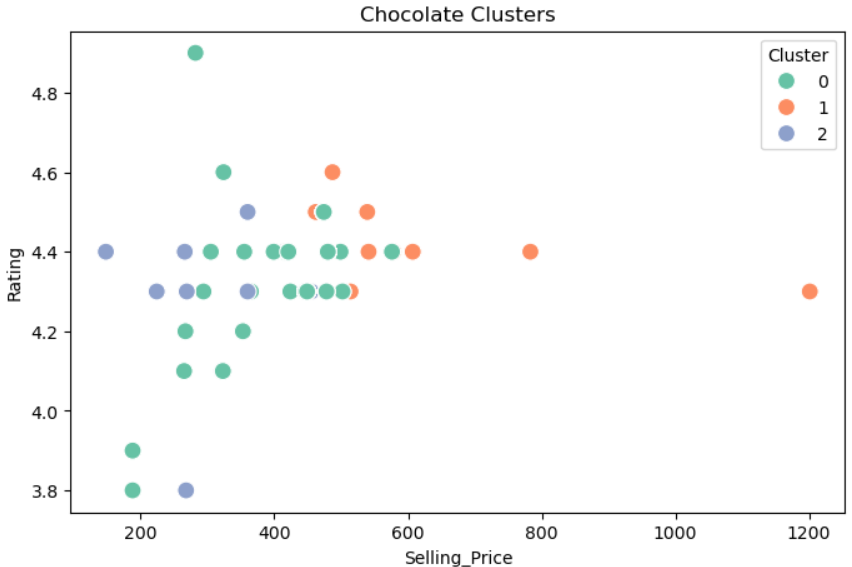
* **Cluster 1:** Budget chocolates (low price, high discount)
* **Cluster 2:** Mid-range chocolates (balanced price and rating)
* **Cluster 3:** Premium chocolates (high rating, low discount, high price)

Used **Elbow Method** to determine optimal cluster count.



#### ****2. Hierarchical Clustering:****

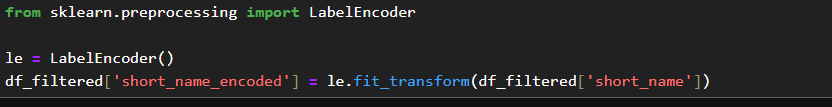
* A **dendrogram** was plotted to visualize how chocolates merge step-by-step into larger clusters.
* The analysis confirmed the presence of **three distinct groups**, aligning with the K-Means results.



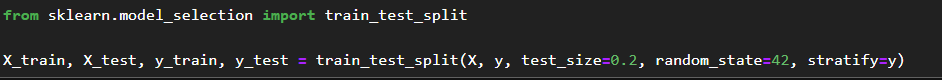
* Chocolates from similar price and rating ranges were grouped together naturally.
* Premium brands like Cadbury Silk or Ferrero Rocher appeared in the same branch of the dendrogram, showing strong feature similarity.
* Budget chocolates (like local or store brands) clustered separately due to lower ratings and higher discounts.

# 11. DATA PREPROCESSING:

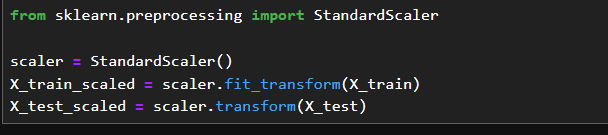
**Label Encoding:** For categorical variables like Brand and Type.



**Train-Test Split:** 80% for training and 20% for testing.

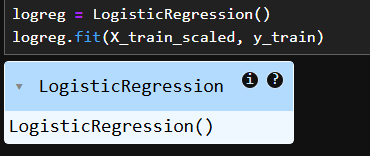


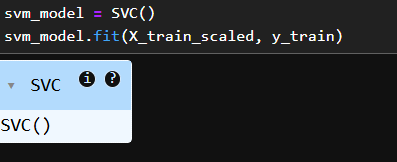
**StandardScaler:** For numeric features such as Price and Rating.

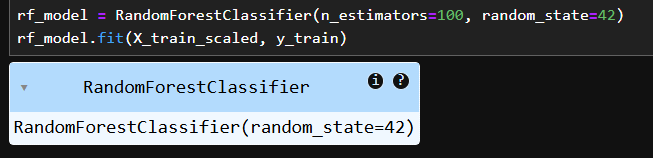


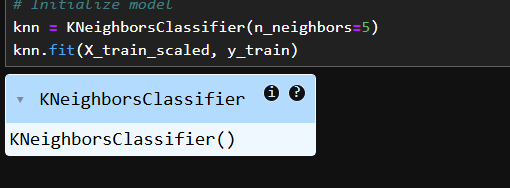
# 12. MODEL TRAINING:

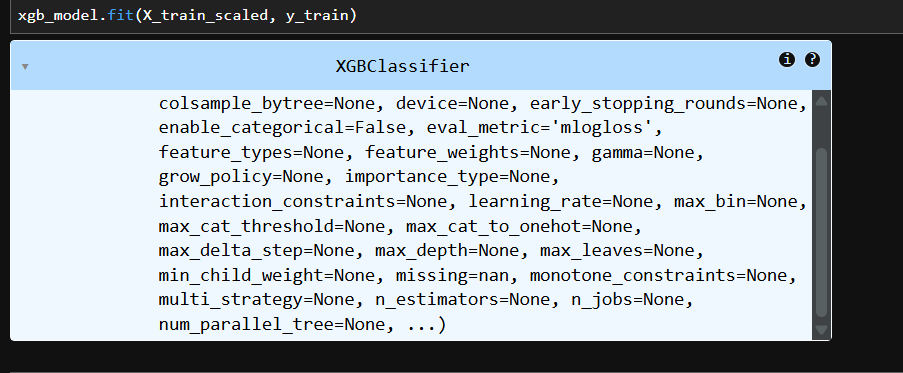
Trained multiple classification models:  
- Logistic Regression  
- Decision Tree  
- Random Forest  
- K-Nearest Neighbors (KNN)  
- Support Vector Machine (SVM)  
- XGBoost





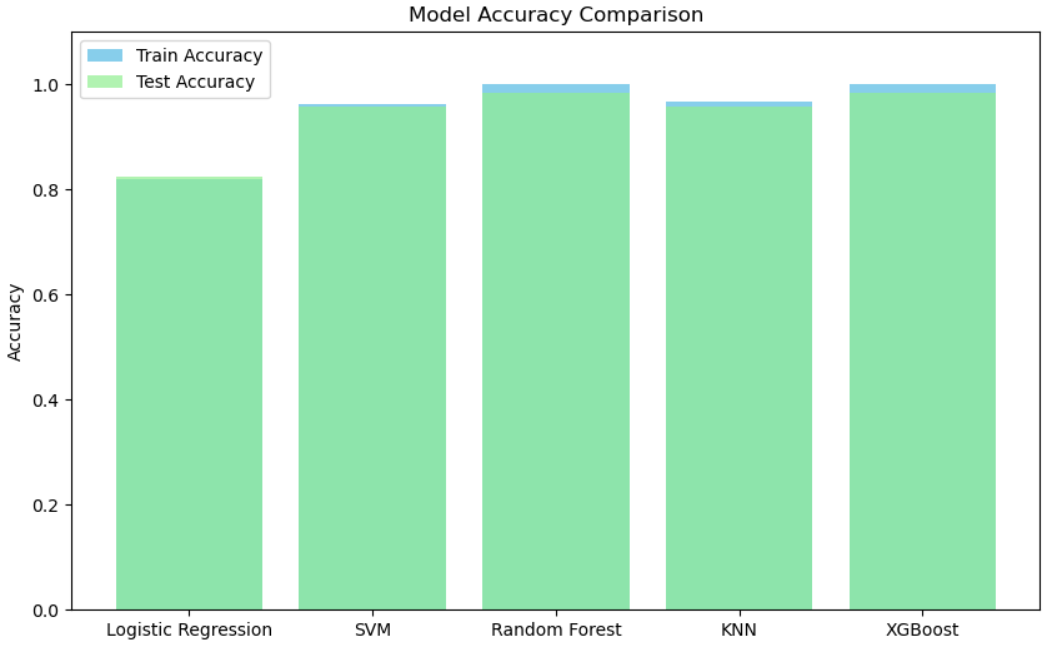






\

# MODEL EVALUATION:

Evaluation Metrics: Accuracy.  
  
Example Results:  


Best Model: XGBoost , KNN, Random Forest achieved the highest accuracy and generalization.

# 

# 14. OVERALL INSIGHTS:

- Price and Discount are strong indicators of chocolate brand and popularity.

- Premium chocolates receive consistently high ratings.

- Discount-heavy chocolates often belong to mid-tier brands

# 15. CONCLUSION:

This project successfully predicted chocolate brand categories using data collected via web scraping. After testing multiple classification models, XGBoost performed best with the highest accuracy. The workflow demonstrates the effective use of data collection, preprocessing, and machine learning for e-commerce analytics.

Future Scope:  
- Expand dataset with **user reviews and taste descriptions**.  
- Deploy as a **Streamlit web app** for live brand prediction.  
- Integrate **NLP** to analyze review sentiments.