MACHINE LEARNING PROJECT

106122112 : SHALU KUMARI

106122034 : DEEPAK KUMAR

TOPIC->

Scraping laptop data from Amazon

ALGORITHM:-

This algorithm describes the complete logic for scraping laptop data from Amazon. It includes fetching data from the website, extracting product information, cleaning and structuring the data, and finally saving it for further analysis. This process helps in gathering product insights like pricing, ratings, and product availability for market research and price comparison.

Execute this project:

- Setup The development environment by installing Python.
- Identify the target Amazon laptop product URL that we want to scrape.
- Develop the scraping code using Python, extract the required data, and store it in a structured format like a CSV file
- Run The code and resolve any bugs within it, like issues with handling dynamic content or working with website change

STEPS→

1. Web Scraping and Data Collection

a) **Send HTTP Requests:**

- ○ Use libraries like `requests` to send HTTP requests to Amazon product URLs.
- ○ Retrieve the HTML content of the web pages.

b) **Parse HTML Content:**

- ○ Use `BeautifulSoup` to parse the HTML content.
- ○ Extract relevant fields such as laptop models, features, and pricing.

c) **Handle Dynamic Content (if applicable):**

- ○ Use tools like `Selenium` to interact with dynamic web elements that load content via JavaScript.

d) **Store Data:**

- ○ Save the extracted data in a structured format such as CSV or JSON files.

2. Data Preprocessing

a). **Data Cleaning:**

- ○ **Remove Duplicates:** Eliminate duplicate entries to ensure the dataset's integrity.
- ○ **Handle Missing Values:** Fill in missing values using techniques like mean imputation or remove incomplete records.

b) **Data Transformation:**

- ○ **Standardize Formats:** Convert all values to a consistent format, such as converting prices to numerical values.
- ○ **Feature Extraction:** Extract relevant features from raw data (e.g., features of laptops).

c) **Data Normalization:** Normalize numerical features (e.g., prices) to a common scale, which improves the performance of ML algorithms.

3. Machine Learning Analysis

a) **Feature Engineering:**

- **Feature Extraction:** Identify and extract key features from the dataset (e.g., laptop specifications, price).
- **Feature Creation:** Create new features that may provide additional insights (e.g., price per feature).

b) **Model Selection:**

- **Choose Algorithms:** Select appropriate machine learning algorithms based on the problem at hand.
  - **Regression Models:** For predicting numerical values (e.g., predicting laptop prices).
  - **Classification Models:** For categorizing data (e.g., classifying laptops into different categories based on features).

c) **Training and Testing:**

- **Split Data:** Divide the dataset into training and testing sets. ○ **Train Model:** Use the training set to train the selected machine learning model.
- **Evaluate Model:** Test the model using the testing set and evaluate its performance using metrics such as accuracy, precision, recall, and F1-score.

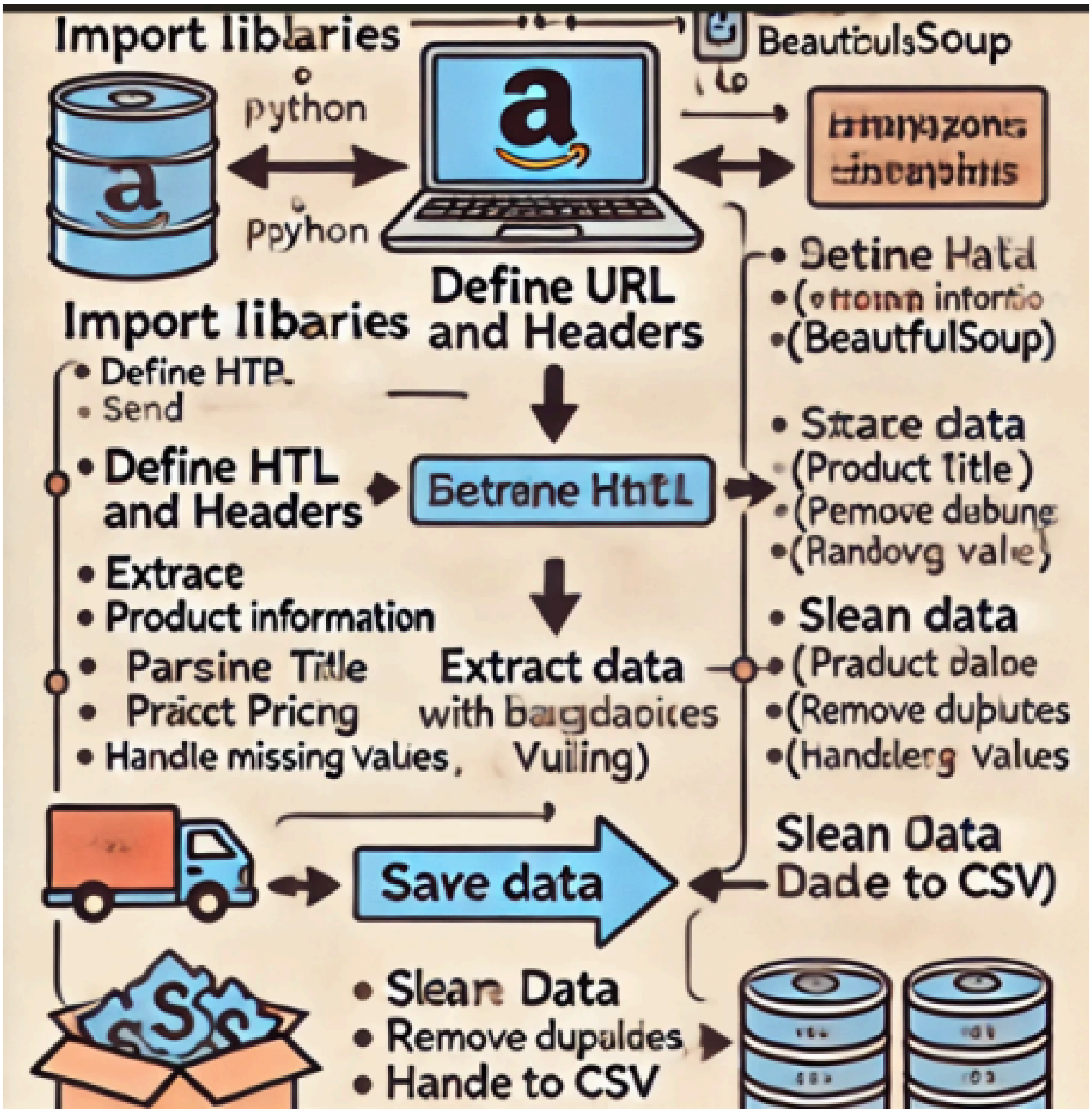
d) Model Deployment:

- **Integration:** Implement the trained model into the scraping workflow to make predictions or analyze new data.
- **Monitoring:** Continuously monitor and update the model as needed to ensure it remains accurate and relevant.

4. Evaluation and Reporting

- **Evaluation Metrics:** Metrics used to assess the performance of machine learning models. For regression tasks, common metrics include Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). For classification tasks, metrics include accuracy, precision, recall, and F1-score.

- **Visualization:** The graphical representation of data to make complex information more accessible and understandable.



→The flowchart illustrates the algorithm for scraping laptop data from Amazon.