**Members: Leader:**

* Martin Mora Victor Bordiga
* Randolfo Castillo **Mentor:**
* Andry Becerra Pedro Dal Bianco
* Gilberto Parra

**Automated product categorization for e-commerce with AI**



**Introduction**

Our project aims to develop a machine-learning model that automatically classifies products across diverse categories, spanning electronics, household items, and beyond, using natural language processing. The primary goal is to create a highly accurate model requiring minimal human intervention. Once completed, the model will be deployed via an API, facilitating efficient and precise product categorization for users.

**Data Collection**

The dataset was obtained from an open repository hosted on GitHub, accessible via the URL 'https://raw.githubusercontent.com/anyoneai/e-commerce-open-data-set/master/products.json'. The data collection process involved accessing this repository, specifically targeting the 'products.json' file containing information about a diverse range of products. This dataset was chosen due to its comprehensive nature, encompassing over 50,000 entries with detailed attributes crucial for retail and e-commerce operations.

The data represents real-world scenarios, making it suitable for training and evaluating machine learning models for product categorization. Upon retrieval, the dataset underwent preprocessing to ensure its readiness for subsequent analysis and model development.

**Data**

Dataset Description:

The dataset comprises comprehensive information on various products essential for retail and e-commerce operations. Each entry is identified by a unique Stock Keeping Unit (SKU) and includes the following attributes:

* Product name
* Type
* Price
* Universal Product Code (UPC)
* Category (presented as a nested dictionary for comprehensive categorization)
* Shipping price.
* Description
* Manufacturer
* Model number
* URL link to the product page
* Image link

The dataset consists of 51,646 entries with 12 columns, each providing valuable information for retail and e-commerce operations, offering a rich source for analyses including pricing trends, product categorization, manufacturer distribution, and other insights crucial for the retail sector.

**Columns description**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| sku | int64 | Stock Keeping Unit, a unique identifier for products. |
| name | object | Name of the product. |
| type | object | Type of the product (e.g., HardGood). |
| price | float64 | Price of the product. |
| upc | int64 | Universal Product Code, a barcode identifier. |
| category | object | A nested dictionary containing information about the product category goes from 1 category per product to six subcategories. |
| shipping | object | Shipping price. |
| description | object | Description of the product. |
| manufacturer | object | Manufacturer of the product. |
| model | object | Model number of the product. |
| url | object | URL link to the product page. |
| image | object | URL link to the product image. |

**Exploratory Data Analysis**

A screenshot of a computer screen

Description automatically generated

The dataset comprises 51,646 entries across 12 columns, representing a diverse range of products essential for retail and e-commerce operations. This initial exploration aimed to gain insights into the dataset's characteristics, uncover patterns, and identify potential challenges or opportunities for model development.

**Missing Values**

There are missing values present in the dataset, particularly in the 'name,' 'manufacturer,' and 'model' columns. These missing values will require handling strategies such as imputation or removal before model training.

**Category Analysis**

**A graph with different colored bars

Description automatically generated**

**Category Distribution Analysis:**

The distribution of products across hierarchical levels of categorization reveals varying degrees of granularity and coverage within each category level.

**Key Insights**

Parent Category: All products are categorized at the parent level, resulting in 100% coverage across top-level categories. This indicates a comprehensive representation of broad product categories within the dataset.

Subcategory 1: Approximately 98.54% of products have a classification at the first subcategory level, suggesting a high level of granularity in the initial categorization.

Subcategory 2: The coverage decreases to approximately 87.14% at the second subcategory level, indicating a reduction in granularity as the categorization becomes more specific.

Subcategory 3: Roughly 50.65% of products are classified at the third subcategory level, representing a significant decrease in coverage compared to previous levels.

Subcategories 4-6: The coverage continues to decline substantially in deeper levels of subcategorization, with only a fraction of products categorized beyond the third level.

*We decided that, because 99.33% of the records are within the first 5 categories, we will use these records to develop and train the model.*

**Here is what the final dataset looks like:**

A screen shot of a computer

Description automatically generated

**Number of name categories per subcategory:**

A graph of different colored rectangular shapes

Description automatically generated

