**Ecommerce Product Categorization**

9. image:

- Description: URL linking to the main image of the product.

- Purpose: Provides visual representation for the product.

10. is\_FK\_Advantage\_product:

- Description: A boolean indicator (True/False) showin**Problem Statement:**

Welcome to KnowledgeHut AI hackathon – **Ecommerce Product Categorization.**

In the rapidly evolving world of eCommerce, accurate product categorization is crucial for ensuring seamless customer experiences, reducing search friction, and increasing product discoverability. However, the sheer volume of diverse products poses a significant challenge. Current classification systems struggle to handle ambiguities, unconventional naming conventions, and multi-language data. This hackathon aims to address these challenges by inviting participants to create innovative solutions that enhance product categorization efficiency, accuracy, and scalability.

Develop a multi-class text classifier that categorizes products with maximum accuracy based on the given dataset.

**Task:**

Your task is to build this model based on the details in this document and submit it. Please read the details carefully before attempting this hackathon.

You will need to decide the following:

1. Analyze the provided dataset to understand key features, missing data, and ambiguities.
2. Do you plan to visualize the data? If so, what kind of plots since its text data?
3. Do you intend to perform data preprocessing tasks such as outlier detection, missing value handling, or feature selection before training your model.?
4. What are your plans to convert text data to numerical or in other terms feature engineering?
5. What is your intended data split ratio for training, validation, and test sets? How do you plan to ensure randomness in this split?
6. Do you anticipate class imbalance in the category feature? If so, how will you address this imbalance?
7. Do you plan to build multi-class text classifier using various ML algorithms?
8. What are your thoughts of using deep learning algorithms?
9. Do you plan to optimize the model by tuning hyperparameters and selecting the most relevant features.?

Your code should have the following:

**Dataset:**

**train\_product\_data.csv: test\_data.csv**

You will note that in this dataset, there are 15,000 records and 15 features.

Here's a detailed description for each column name:

1. uniq\_id:

- Description: A unique identifier for each product.

- Purpose: Acts as the primary key to distinguish each product record uniquely.

2. crawl\_timestamp:

- Description: The timestamp when the product data was last scraped or collected.

- Purpose: Helps identify the data's recency and track changes over time.

3. product\_url:

- Description: The URL linking directly to the product's page on the eCommerce platform.

- Purpose: Allows direct access to the product's information and purchasing page.

4. product\_name:

- Description: The name or title of the product as displayed on the eCommerce platform.

- Purpose: Provides a searchable and readable identification of the product.

5. **product\_category\_tree:**

- Description: The hierarchical structure representing the product's category on the platform.

- Purpose: Useful for categorization, analysis, and filtering of products.

6. pid:

- Description: A unique identifier specific to the eCommerce platform for each product.

- Purpose: Used to reference products internally on the platform.

7. retail\_price:

- Description: The original or retail price of the product before any discounts.

- Purpose: Helps understand the product's standard market value.

8. discounted\_price:

- Description: The price of the product after applying any discounts or offers.

- Purpose: Reflects the final price a customer would pay.

g if the product is part of Advantage program.

- Purpose: Denotes if the product has additional benefits like faster delivery or special quality checks.

11. **description**:

- Description: Detailed information about the product, including features, specifications, and usage.

- Purpose: Helps customers understand the product's value proposition and unique selling points.

12. product\_rating:

- Description: The product's overall rating on the platform, based on customer reviews.

- Purpose: Indicates customer satisfaction and product quality.

13. overall\_rating:

- Description: The aggregate rating of the product across different platforms or periods.

- Purpose: Offers a comprehensive view of the product's reception.

14. brand:

- Description: The name of the brand or manufacturer of the product.

- Purpose: Assists in brand-based analysis and filtering.

15. product\_specifications:

- Description: Detailed specifications of the product, often in JSON or structured format.

- Purpose: Provides technical and functional details to aid customer decision-making.

**test\_data.csv**:

This dataset can be used to test your model, in this dataset there are ~2,500 records and 14 features similar as in the train data (excluding the target ‘**product\_category\_tree**’)

**test\_results.csv:**

This dataset contains the target (‘**product\_category\_tree**’ of the ~2500 test data) and can used to evaluate your model performance.

**Submission:**

Please submit the Jupyter Notebook. This should clearly show:

1. The code you have written

2. The output of the test data

3. A comprehensive description of each code block, together with the decisions you've made and the rationale for those decisions.

4. A comprehensive description of what you also tried that did not work, and what lessons you have learnt from this hackathon.

5. Powerpoint presentation on your work

6. Video Walkthrough

7. A GitHub link with all the above deliverables

8. Streamlit app (Optional)

**Note:**

﻿• Use Python programming language

• Use a Google Colab, or a standard laptop/desktop to build the model

• This is a challenging dataset to predict accurately on, so iterate your approach over time making note of what works and what doesn't. In the long term, this is as useful to you as getting a high model accuracy.

• Avoid Plagiarism as the objective of this exercise is to give you a real-world project to build and we hope you will use this opportunity wisely to your benefit.