**Data Science Hackathon**

**Hackathon Topic :– Ecommerce Product Categorization**

**Introduction:** Welcome to the "**Ecommerce Product Categorization**" hackathon, hosted by KnowledgeHut! This exciting competition challenges you to apply your data science, machine learning & NLP skills to a real-world problem in the retail sector. Get ready to demonstrate your expertise in predictive modeling and NLP while addressing the crucial time-consuming issue of product categorization.

**Background:** In this hackathon, you'll have the opportunity to showcase your proficiency in data analysis, Python programming, statistical analysis, machine learning, natural language processing, deep learning and model evaluation. The primary focus of this project is to assess your ability to build and train a multi-class text classifier that can predict the category of the product using description.

**Hackathon Tasks:** Your mission in this hackathon is to build and train a multi-class text classifier model using the dataset provided. Your task includes the following:

**1. Analyze the Dataset:**

- Explore the provided dataset to identify key features, missing data, and ambiguities.

**2. Visualize the Data:**

- Generate meaningful visualizations to uncover data patterns and correlations, considering the text-based nature of the dataset.

**3. Preprocess the Data:**

- Clean and transform the dataset by handling missing values, outliers, and inconsistencies.

**4. Feature Engineering/Text to Features:**

- Convert raw product data (text descriptions, specifications) into informative features using techniques like TF-IDF or word embeddings.

**5. Address Class Imbalance:**

- Detect and address class imbalance in the product category feature using sampling techniques or class weights.

**6. Develop Machine Learning/Deep Learning Models - Multi-class text classifier:**

- Design and train predictive models, considering both machine learning and deep learning algorithms for product categorization.

**7. Evaluate Model Accuracy and Other Metrics:**

- Measure model performance using evaluation metrics like accuracy, F1 score, precision, and recall.

**8. Fine-Tune Models:**

- Optimize model performance by tuning hyperparameters and selecting the most relevant features.

**9. Predict on Test Data:**

- Evaluate the final model on the test dataset to validate its performance.

**10. Create a GitHub Repository:**

- Share all code, notebooks, and documentation in a structured format through a public GitHub repository.

**11. Prepare a PowerPoint Presentation:**

- Summarize findings and methodology in a concise PowerPoint presentation.

**12. Create a Video Walkthrough:**

- Record a comprehensive video walkthrough explaining the approach, methodology, and results.

**13. Build a Streamlit App (Optional):**

- Develop an interactive web app using Streamlit to showcase model predictions and visualizations.

**Key Objectives:** In this hackathon, we expect you to accomplish the following key objectives:

**1. Data Exploration and Preparation:**

- Explore and analyze the dataset to understand key features, detect missing data, and identify ambiguities.

- Preprocess the dataset for consistency, noise reduction, and missing value handling.

**2. Descriptive Analysis:**

- Perform descriptive analysis to identify data patterns, category distributions, and inconsistencies.

- Visualize data insights to guide feature engineering and model development.

**3. Feature Engineering/Text to Features:**

- Transform raw product data into informative features suitable for machine learning models.

- Convert text descriptions into numerical features using techniques such as TF-IDF, word embeddings, or custom methods.

**4. Predictive Modeling – Multi class text classifier:**

- Design and develop multi-class text classifier for accurate product categorization.

- Consider classification, clustering, or hybrid models based on problem requirements.

- Ensure the model can handle ambiguous products

**5. Fine Tuning:**

- Optimize models by tuning hyperparameters and selecting relevant features.

- Evaluate and compare models using metrics such as accuracy, F1 score, precision, and recall.

- Validate models through cross-validation or a separate test dataset.

**6. Enhance Categorization Accuracy:**

- Improve model accuracy by incorporating domain-specific knowledge or ensemble methods.

- Ensure solutions can accurately categorize unconventional product names and handle ambiguities.

- Develop models that support scalable, real-time categorization.

**Why Participate:** Participating in the hackathon offers several compelling reasons to join:

* Showcase your expertise in data analysis, machine learning, and predictive modeling in the context of retail text data.
* Apply your skills to real-world data
* Collaborate with like-minded enthusiasts in a competitive and immersive environment, fostering valuable connections.
* Enhance your portfolio with a retail-focused project that demonstrates your data science capabilities.
* Compete for exciting prizes and recognition, showcasing your skills as a data scientist and model builder.

We look forward to your participation and the insights you will bring to this challenge!