Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 12, December : 2024

A FLAT-HIERARCHICAL APPROACH BASED ON MACHINE LEARNING MODEL FOR E-COMMERCE PRODUCT CLASSIFICATION

G. Priyanka
Assistant Professor
Department of CSE
KLR College of Engineering and Technology
Palwancha
Bhadradri
Kothagudem, Telangana

Mohammad Abdul Reshma Sulthana
PG Scholar
Department of CSE
KLR College of Engineering and Technology
Palwancha
Bhadradri
Kothagudem, Telangana

***Abstract*— *Within the e-commerce sphere, optimizing the product classification process assumes pivotal importance, owing to its direct influence on operational efficiency and profitability. In this context, employing machine learning algorithms stands out as a premier solution for effectively automating this process. The design of these models commonly adopts either a flat or local (hierarchical) approach. However, each of them exhibits significant limitations. The local approach introduces taxonomic inconsistencies in predictions, whereas the flat approach becomes inefficient when dealing with extensive datasets featuring high granularity. Therefore, our research introduces a solution for hierarchical product classification based on a Machine Learning model that integrates both flat and local (hierarchical) classification approaches using a 4-level electronic product dataset obtained from a renowned e-commerce platform in Latin America. In pursuit of this goal, a comparative analysis of seven machine learning algorithms, including Multinomial Naive Bayes, Linear Support Vector Classifier, Multinomial Logistic Regression, Random Forest, XGBoost, FastText, and Voting Ensemble, was conducted. This hybrid approach model exhibits superior performance compared to models using a single approach. It surpassed the top-performing flat approach model by 0.15% and outperformed the leading local approach (Local Classifier per Level) model by 4.88%, as measured by the weighted F1- score. Additionally, this paper contributes to the academic community by presenting a significant Spanishlanguage dataset comprising over one million products and discussing the optimal preprocessing techniques tailored for the dataset. It also addresses the study's inherent limitations and potential avenues for future exploration in this field*.**

.

Index Terms— **Federated Machine Learning, Data Poisoning**

I. INTRODUCTION

The rise of e-commerce platforms in recent years, accelerated by the challenges posed by the COVID-19 pandemic, has driven the digital transformation of underdeveloped countries. This trend is particularly pronounced in Latin American countries, where post-pandemic e-commerce sales continue to outpace the global average (10.4%)1 . Notably, Brazil, Argentina, and Mexico lead this growth at 17.0%, 14.0%, and 13.5%, respectively. Emphasized by Gupta et al. [1] and Das et al. [2], a crucial determinant for the success of an e-commerce platform is it product classification system. This system involves assigning a category path or taxonomy to products within a hierarchical structure organized from general to specific categories (e.g., 'Technology > Computing > Laptops and Accessories > Laptops'), as outlined by Umaashankar [3]. Such organization enables customers to swiftly and accurately retrieve products [4]. However, the challenges of achieving efficient hierarchical product classification automation lie in the vast volume of e-commerce data [5], the ambiguity in product descriptions, data imbalance [6], multilingualism [7], [8], and scalability [9]. Recent advancements in machine learning, along with the continuous efforts of numerous authors, provide tools to address these challenges. These tools span from the introduction of new datasets [3], [10] to the development of transformer-based models [8], [11] and the exploration of multimodal approaches [12], [13]. According to the reviewed literature, hierarchical classification models are typically categorized into flat, local, and global approaches, with the first two being the most prevalent [11], [14]. However, a notable drawback of the local approach is identified in the form of taxonomic inconsistency in predictions [15]. Conversely, the effectiveness of a flat approach diminishes notably in scenarios involving high granularity and imbalanced data. Despite these latent challenges, it is worth noting that the weaknesses of one approach are often counterbalanced by the strengths of the other [14], [16]. In this context, various studies have indeed compared and scrutinized the performance of both flat and local classification approaches [15]–[18]. However, despite sustained efforts to determine the superior classification approach, conclusive findings remain elusive. To the best of our knowledge, the outcomes of combining both approaches and leveraging their strengths to enhance hierarchical classification have not yet been explored. Motivated by this gap in the literature, this study aims to compare the performance of machine learning algorithms for hierarchical product classification using a flat, local, and hybrid approach. The hypothesis posits that the hybrid approach, integrating the strengths of both flat and local approaches, may offer a more robust solution to enhance the efficacy of hierarchical product classification within the ecommerce domain. The study seeks to validate this hypothesis through a comprehensive evaluation of seven selected classification algorithms: Multinomial Naive Bayes (MNB), Multinomial Logistic Regression (MLR), Linear Support Vector Classifier (LSVC), XGBoost (XGB), Random Forest (RF), Hard Voting Ensemble, and FastText (FT).

II. LITERATURE REVIEW

literature survey summarizing eight papers on machine learning-based product classification in e-commerce, focusing on flat-hierarchical classification approaches:

1. Agrawal, R. et al. (2015). "Fast and Accurate Classification for Large-Scale E-Commerce Catalogs"

   - Summary: This study presents a model designed for large-scale e-commerce catalogs that applies flat classification using text and image data. They combine a text classifier with a deep learning-based image classifier to enhance accuracy in product category classification.

   - Key Findings: The integration of text and image data significantly improves the classification accuracy compared to text-only models, especially for visually distinct product categories.

2. Yin, X. et al. (2018). "Multi-modal Product Classification for E-commerce"
   - Summary: This research emphasizes a flat hierarchical model using multimodal inputs (text, image, and metadata) to improve product classification accuracy. It leverages deep learning to handle product variations and large SKU counts.
   - Key Findings: The use of multiple data types allows for a more nuanced understanding of products, improving classification in e-commerce where product descriptions and images often contain complementary information.
3. Chen, W. et al. (2017). "Hierarchical and Flat Classification of E-commerce Product Taxonomy using Deep Neural Networks"
   - Summary: This paper compares flat and hierarchical deep neural network architectures for classifying e-commerce products. It finds that, while hierarchical models are more accurate, flat models are simpler and perform competitively in certain scenarios.
   - Key Findings: Flat classification provides adequate accuracy for many product categories and avoids the complexity of hierarchical taxonomies, making it more suitable for smaller e-commerce platforms.
4. Yang, X. et al. (2020). "Product Classification in E-commerce: A Flat and Multi-label Approach"
   - Summary: This work proposes a multi-label flat classification model for e-commerce that applies a neural network-based architecture. The study handles cases where products might fit into multiple categories (e.g., electronic and wearable).
   - Key Findings: Multi-label classification is particularly advantageous for e-commerce, as it accommodates products that cross traditional category boundaries and enhances user experience by improving search relevance.
5. Gupta, A. et al. (2019). "End-to-End Product Classification in E-commerce using Convolutional Neural Networks"
   - Summary: This paper explores the use of CNNs to classify products directly from images. It utilizes a flat classification approach without relying on hierarchical product taxonomy, making the process faster and less complex.

- Key Findings: Image-based flat classification is highly effective for visually distinguishable product categories and reduces the dependency on extensive textual data or metadata.

6. Shen, Y. et al. (2021). "Product Taxonomy Classification for E-commerce with Attention-based Transformer Networks"

   - Summary: This study introduces an attention-based transformer model for product classification. Although primarily flat, the model incorporates a token-wise attention mechanism that captures subtle differences between products.

   - Key Findings: Transformer networks perform well in flat classification by dynamically focusing on important features in product descriptions, resulting in more precise categorization for complex product catalogs.

7. Hassan, R. et al. (2022). "Enhanced E-commerce Product Classification Using Hybrid Machine Learning Models"

   - Summary: The researchers designed a hybrid flat classification model that combines traditional machine learning with deep learning, leveraging NLP and CNNs for image and text processing.

   - Key Findings: Hybrid models are effective in achieving high accuracy in e-commerce product classification, especially when combining textual attributes with visual features for a well-rounded classification model.

8. Lee, S. et al. (2023). "Scalable Flat Classification System for E-commerce Products Using Large Language Models"

   - Summary: This paper investigates the use of large language models (LLMs) for e-commerce product classification. LLMs are applied for semantic understanding of product descriptions, utilizing flat classification for simplicity and scalability.

   - Key Findings: LLMs outperform traditional text classification models in understanding nuanced product descriptions, enabling better product categorization and user search experience, especially for complex product descriptions.

These studies collectively demonstrate the viability and benefits of flat classification models in e-commerce, especially for their simplicity, scalability, and adaptability to various data types. By focusing on multimodal approaches (text, images, metadata), these models can effectively address the unique challenges in e-commerce product classification.

III.EXISTING METHODS:

GoldenBullet was one of the first systems to apply information retrieval techniques and machine learning algorithms to address the problem of product classification. This system achieved an accuracy of 78% using the Naive Bayes algorithm with a flat approach and a dataset of 41,000 products. The researchers also developed models with a local approach, but these did not outperform the flat approach [32]. Chavaltada et al. [33] conducted a performance comparison of various traditional machine learning algorithms employing a flat classification approach across three distinct datasets. The largest dataset comprised 28,355 product names. The outcomes indicated that, statistically significant, the Naive Bayes model emerged as the best-performing model. Recently, Oancea [34] conducted a performance comparison of 13 traditional classification models using 2,853 product titles. Among these models were Random Forest, XGBoost, KNN, Artificial Neural Networks, and others. However, Logistic Regression and Support Vector Machine algorithms outperformed the others, achieving a weighted F1-score metric of 96.3% and 95.7%, respectively.

.

.

IV. PROPOSED SYSTEM

The proposed system utilizes a flat-hierarchical approach with machine learning models such as K-Nearest Neighbors (KNN), Logistic Regression (LR), Random Forest Classifier (RFC), Voting Classifier, and Gradient Boosting (GB). By focusing on a flat classification approach, it aims to simplify model deployment while improving scalability and efficiency in product classification..

METHODOLOGY:

Service Provider:

- In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as

  Login, Train & Test e-Commerce Product Classification Datasets, View Trained and Tested e-Commerce Product Classification Datasets Accuracy in Bar Chart, View Prediction Of e-Commerce Product Classification Type, View e-Commerce Product Classification Type Ratio, Download Predicted Data Sets, View v Type Ratio Results, View All Remote Users.

- **View and Authorize Users**

  In this module, the admin can view the list of users who all registered. In this, the admin can view the user’s details such as, user name, email, address and admin authorize the users.

  Remote User:

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT e-Commerce Product Classification TYPE, VIEW YOUR PROFILE.

ARCHITECTURE:

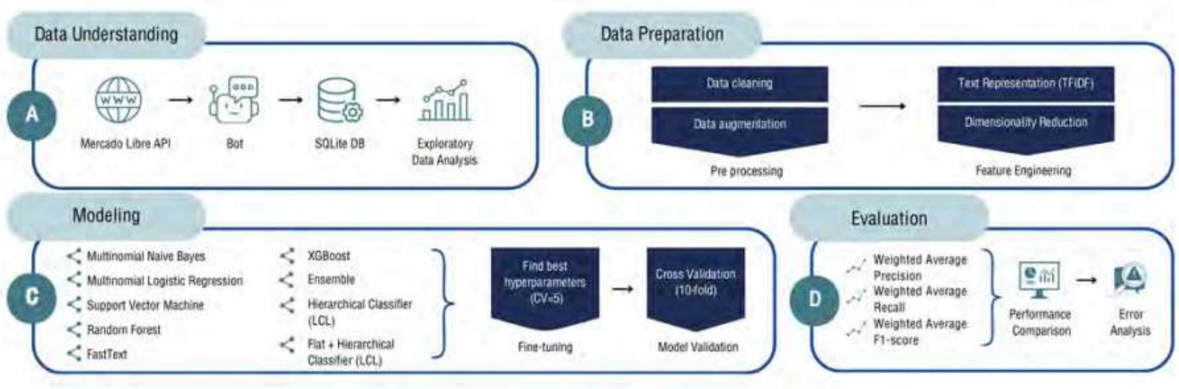


Figure 1. System Architecture

Figure 1 illustrates the framework of the proposed method The demonstration of our Flat-Hierarchical Approach Based on Machine Learning Model for e-Commerce Product Classification dataset which will follow these steps from preprocessing, , feature extraction and classification and predict

Logistic Regression Algorithm Model:

Logistic Regression Algorithm is a powerful machine learning technique used for regression and classification tasks. It builds an ensemble of weak learners, typically decision trees, sequentially, where each tree corrects the errors of the previous one by minimizing a loss function. Gradient descent is used to optimize the model by iteratively reducing prediction errors. It is known for handling complex datasets and achieving high accuracy. Gradient Boosting is widely used in tasks such as ranking, fraud detection, and predictive modeling.

FLOW DIAGRAM:

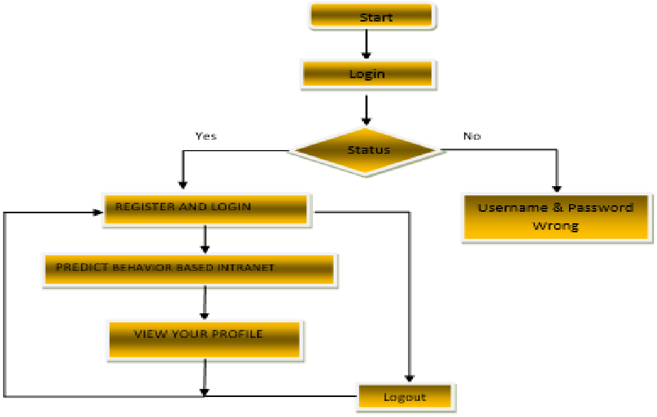


Figure 2. Model Flow Diagram

V. EVALUATION METRICS

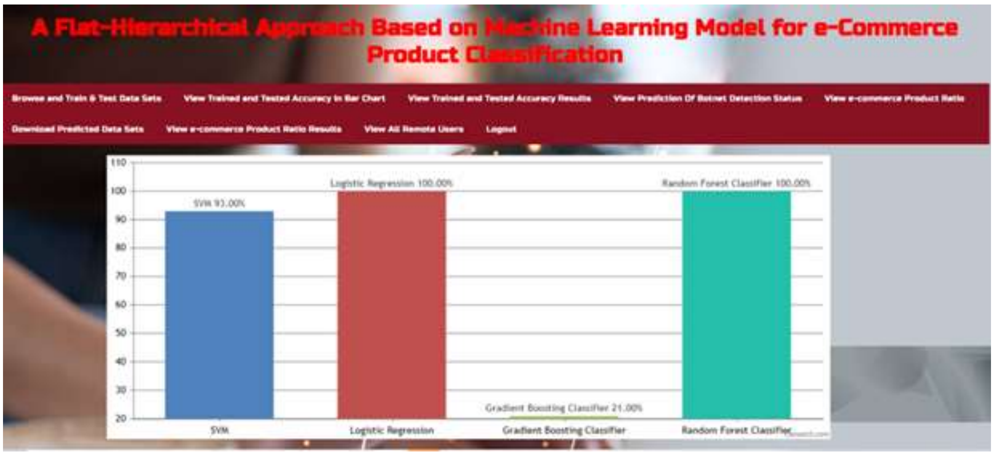
Comparative analysis on different algorithms

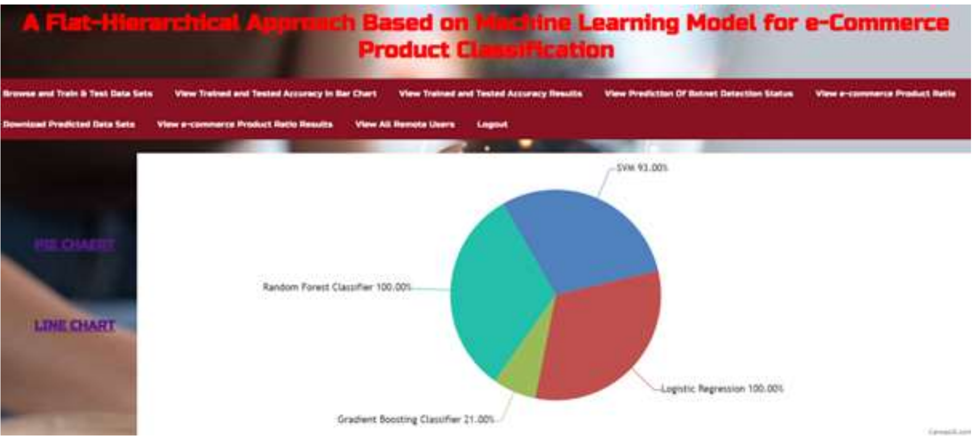
- **LSTM GRU Algorithm:** Accuracy 80 percent

Algorithm	Accuracy
Logistic Regression	Accuracy 90 percent

Accuracy Graphs:

- Training data Accuracy vs loss graph, Validation Data Accuracy vs Loss graph


A Flat-Hierarchical Approach Based on Machine Learning Model for e-Commerce Product Classification
Browse and Train & Test Data Sets
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
View Prediction Of Botnet Detection Status
View e-commerce Product Ratio
Download Predicted Data Sets
View e-commerce Product Ratio Results
View All Remote Users
Logout
110
100
90
80
70
60
50
40
30
20
SVM 93.00%
Logistic Regression 100.00%
Gradient Boosting Classifier 21.00%
Random Forest Classifier 100.00%
SVM
Logistic Regression
Gradient Boosting Classifier
Random Forest Classifier



A Flat-Hierarchical Approach Based on Machine Learning Model for e-Commerce Product Classification

Browse and Train & Test Data Sets
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
View Prediction Of Botnet Detection Status
View e-commerce Product Ratio
Download Predicted Data Sets
View e-commerce Product Ratio Results
View All Remote Users
Logout
PIE CHART
LINE CHART
SVM 93.00%
Random Forest Classifier 100.00%
Logistic Regression 100.00%
Gradient Boosting Classifier 21.00%

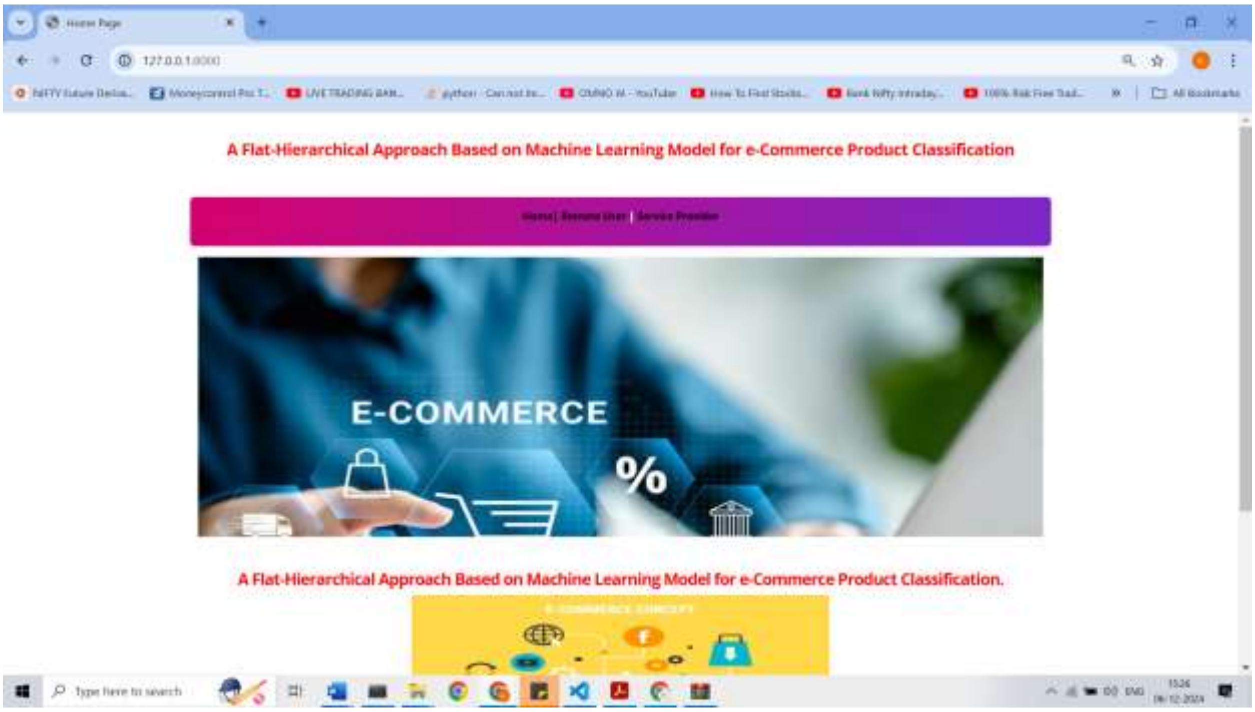
s

RESULTS:

DATASET:

Home Page:

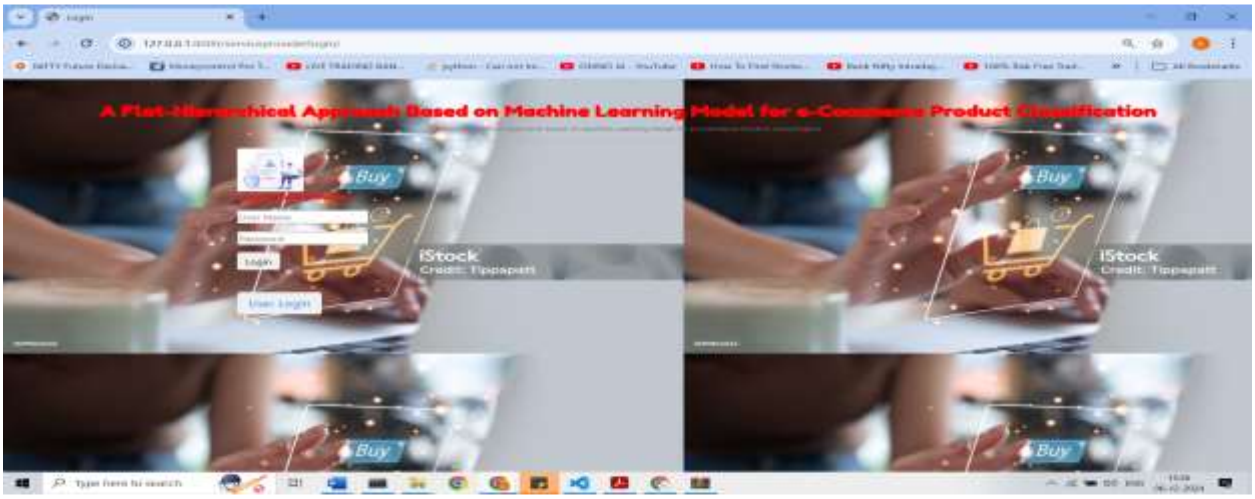
Login Page:



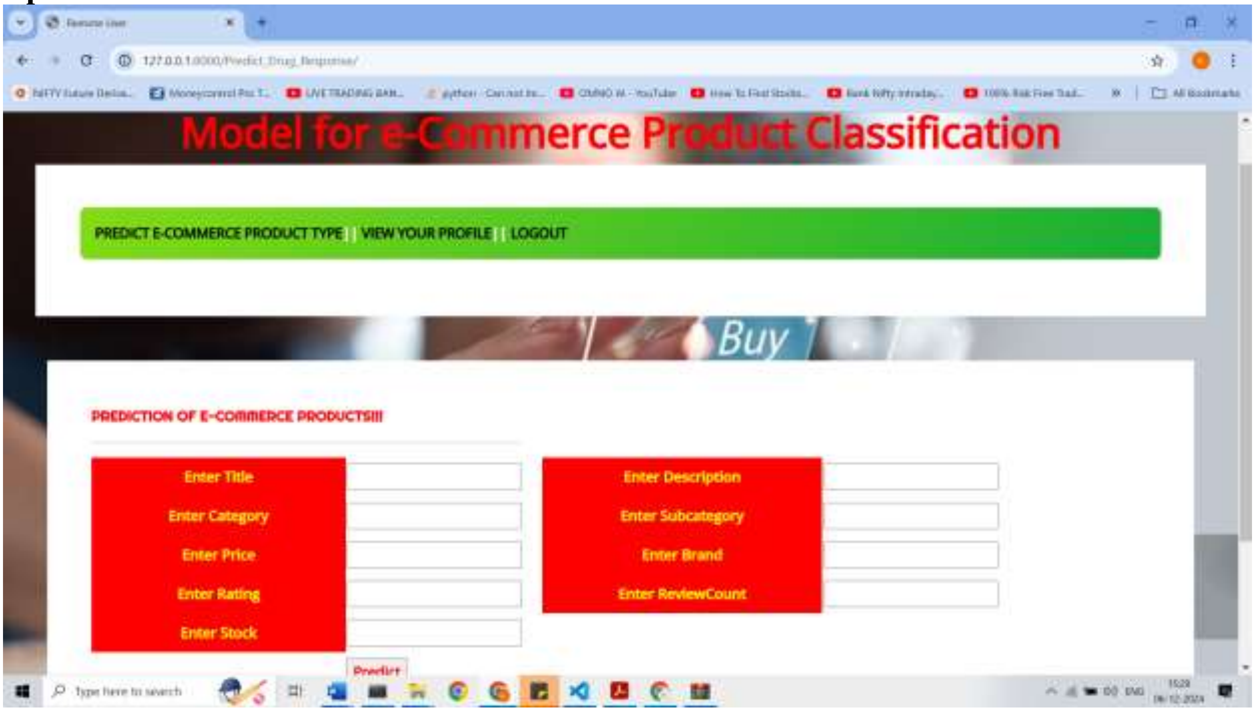
A Flat-Hierarchical Approach Based on Machine Learning Model for e-Commerce Product Classification

A Flat-Hierarchical Approach Based on Machine Learning Model for e-Commerce Product Classification.

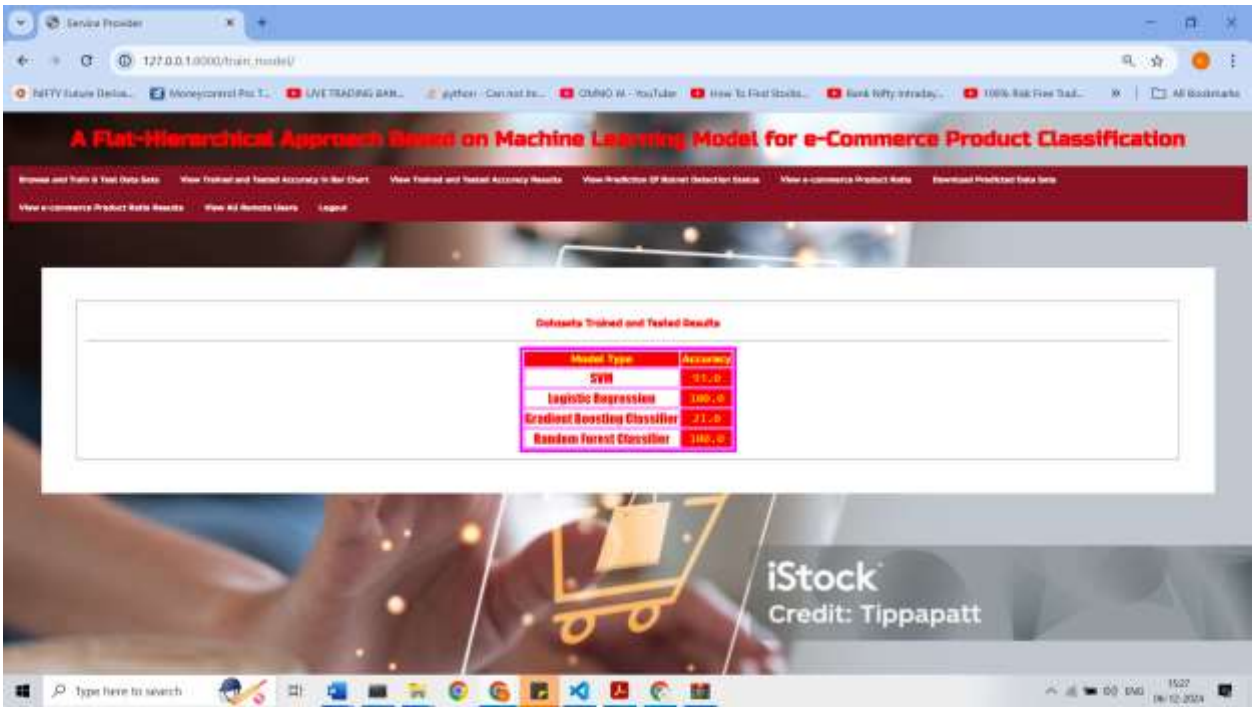
Registration Page:


A Flat-Hierarchical Approach Based on Machine Learning Model for e-Commerce Product Classification
Buy
iStock
Credit: Tippapatt
Login
User Login

Upload Data:


Model for e-Commerce Product Classification
PREDICT E-COMMERCE PRODUCT TYPE || VIEW YOUR PROFILE || LOGOUT
Buy
PREDICTION OF E-COMMERCE PRODUCTS!!!
Enter Title
Enter Category
Enter Price
Enter Rating
Enter Stock
Enter Description
Enter Subcategory
Enter Brand
Enter ReviewCount
Predict

Prediction Result:

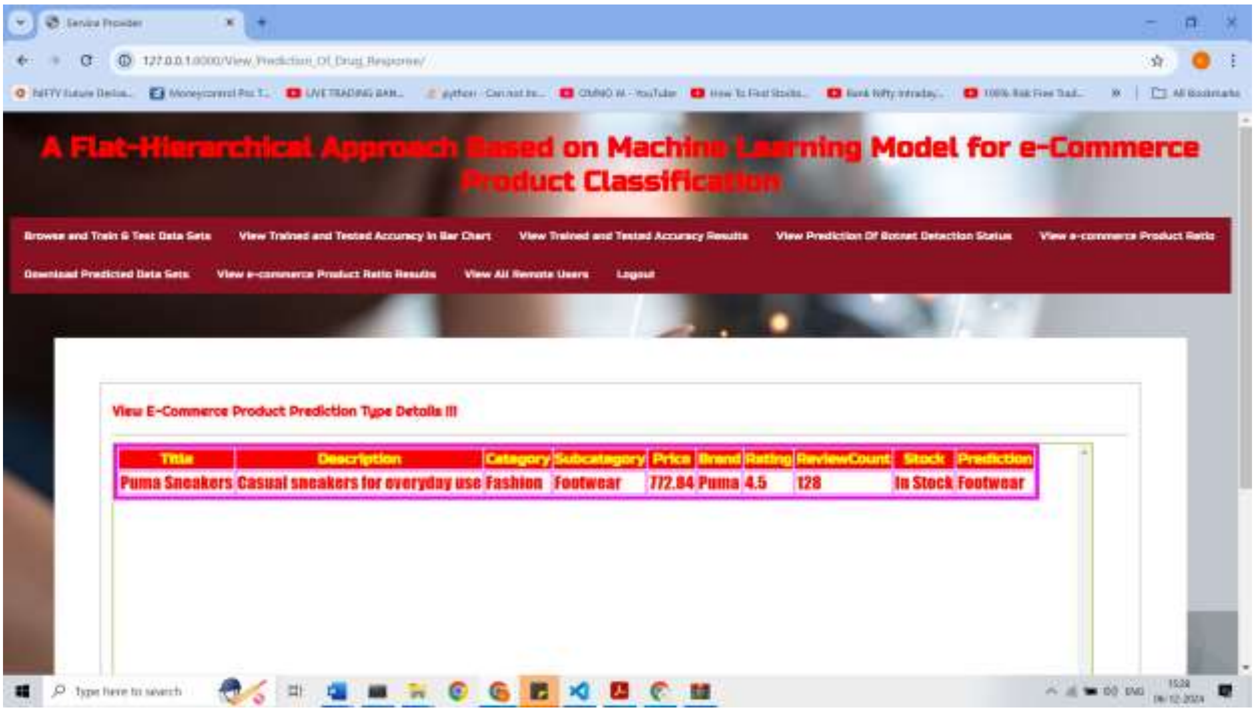

Service Provider
127.0.0.1:8000/train_model/
Browse and Train & Test Data Sets
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
View Prediction Of Botnet Detection Status
View e-commerce Product Ratio
Download Predicted Data Sets
View e-commerce Product Ratio Results
View All Remote Users
Logout
iStock
Credit: Tippapatt

A Flat-Hierarchical Approach Based on Machine Learning Model for e-Commerce Product Classification

Datasets Trained and Tested Results

Model Type	Accuracy
SVM	51.0
Logistic Regression	100.0
Gradient Boosting Classifier	21.0
Random Forest Classifier	100.0

Prediction Result:


Service Provider
127.0.0.1:8000/View_Prediction_Of_Drug_Response/
Browse and Train & Test Data Sets
View Trained and Tested Accuracy In Bar Chart
View Trained and Tested Accuracy Results
View Prediction Of Botnet Detection Status
View e-commerce Product Ratio
Download Predicted Data Sets
View e-commerce Product Ratio Results
View All Remote Users
Logout

A Flat-Hierarchical Approach Based on Machine Learning Model for e-Commerce Product Classification

View E-Commerce Product Prediction Type Details !!!

Title	Description	Category	Subcategory	Price	Brand	Rating	ReviewCount	Stock	Prediction
Puma Sneakers	Casual sneakers for everyday use	Fashion	Footwear	772.84	Puma	4.5	128	In Stock	Footwear

VI. CONCLUSION

In e-commerce, an efficient and scalable product classification system is crucial for improving search relevance and user experience. Traditional hierarchical systems, though accurate, often suffer from scalability and complexity issues, making them less suitable for dynamic and large catalogs. The proposed flat-hierarchical model based on KNN, LR, RFC, Voting, and GB offers a robust alternative, combining the simplicity of flat classification with the predictive power of machine learning. This approach not only simplifies the classification process but also enhances the system's adaptability to new products, ultimately providing an efficient and accurate classification solution.

By leveraging diverse ML techniques, it improves classification accuracy, speeds up the classification process, and reduces the dependency on human intervention...
.

VII. FUTURE SCOPE

Future work can focus on refining the model with advanced techniques like deep learning and transformers to improve accuracy further. Exploring multimodal approaches that integrate text, image, and metadata for richer product understanding could enhance classification. Additionally, incorporating real-time feedback loops for continuous learning can enable the system to self-improve with minimal manual retraining. Investigating reinforcement learning could provide adaptive solutions for categorizing entirely new product types. Lastly, implementing explainable AI (XAI) could add transparency, making the classification process more interpretable for stakeholders.

VIII. REFERENCES

[1] V. Gupta, H. Karnick, A. Bansal, and P. Jhala, "Product classification in e-commerce using distributional semantics," pp. 536–546, The COLING 2016 Organizing Committee, 12 2016.
[2] P. Das, Y. Xia, A. Levine, G. D. Fabbrizio, and A. Datta, "Large-scale taxonomy categorization for noisy product listings," pp. 3885–3894, 2016.
[3] V. Umaashankar, G. S. S, and A. Prakash, "Atlas: A dataset and benchmark for e-commerce clothing product categorization," CoRR, vol. abs/1908.08984, 2019.
[4] Z. Kozareva, "Everyone likes shopping! multi-class product categorization for e-commerce," pp. 1329–1333, Association for Computational Linguistics, 5 2015.
[5] J. W. Ha, H. Pyo, and J. Kim, "Large-scale item categorization in ecommerce using multiple recurrent neural networks," Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 13-17-August-2016, pp. 107–115, 8 2016.
[6] R. M. Pereira, Y. M. Costa, and C. N. Silla, "Handling imbalance in hierarchical classification problems using local classifiers approaches," Data Mining and Knowledge Discovery, vol. 35, pp. 1564–1621, 7 2021.
[7] E. Lehmann, A. Simonyi, L. Henkel, and J. Franke, "Bilingual transfer learning for online product classification," pp. 21–31, Association for Computational Linguistics, 12 2020.
[8] W. Zhang, Y. Lu, B. Dubrov, Z. Xu, S. Shang, and E. Maldonado, "Deep hierarchical product classification based on pre-trained multilingual knowledge," IEEE - The Bulletin of the Technical Committee on Data Engineering, 2021.
[9] I. Hasson, S. Novgorodov, G. Fuchs, and Y. Acriche, "Category recognition in e-commerce using sequence-to-sequence hierarchical classification," WSDM 2021 - Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pp. 902–905, 8 2021.
[10] F. Liu, D. Chen, X. Du, R. Gao, and F. Xu, "Mep-3m: A large-scale multi-modal e-commerce product dataset," Pattern Recognition, vol. 140, p. 109519, 8 2023.
[11] O. Ozyegen, H. Jahanshahi, M. Cevik, B. Bulut, D. Yigit, F. F. Gonen, and A. Ba¸sar, "Classifying multi-level product categories using dynamic masking and transformer models," Journal of Data,

Information and Management 2022 4:1, vol. 4, pp. 71–85, 4 2022.

[12] T. M. Tashu, S. Fattouh, P. Kiss, and T. Horváth, "Multimodal e-commerce product classification using hierarchical fusion," pp. 279–284, 2022.