iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 12 **Issue:** XII **Month of publication:** December 2024

**DOI:** https://doi.org/10.22214/ijraset.2024.65845

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 12 Issue XII Dec 2024- Available at www.ijraset.com*

# Insight Cart- A Product Analysis Website Using Web Scraping and Machine Learning

Abhishek K S[1], Ravindra S[2], Archana B H[3], Abhishek H J[4], Abhishek S Palankar[5]
*Department of Computer Science and Engineering, Jawaharlal Nehru New College of Engineering, Visvesvaraya Technological University- Belagvi, Shivamogga 577204- Karnataka*

***Abstract:*** ***This paper presents an innovative approach to e-commerce data analysis using web scraping and machine learning techniques. By aggregating data from platforms like Amazon, eBay, Snapdeal, and Ajio, the system facilitates product comparison and price prediction. Leveraging state-of-the-art scraping tools such as Selenium and BeautifulSoup, and machine learning algorithms, including K-Means clustering and Linear Regression, the system ranks products and forecasts price trends. A detailed graph showcasing predicted price trends over time provides actionable insights for consumer decision-making. Results emphasize the integration of analytics and data visualization to enhance the e-commerce experience.***
***Keywords: Web Scraping, Machine Learning, Price Prediction, E-Commerce, Data Visualization.***

## I. INTRODUCTION

The e-commerce industry is experiencing exponential growth, presenting consumers with an overwhelming number of products and prices across multiple platforms. Identifying the best time to purchase or comparing similar products often requires considerable effort due to scattered and inconsistent data. This paper addresses these challenges by introducing a system that integrates web scraping and machine learning (ML). Using data from platforms such as Amazon , Ajio, Snapdeal and eBay, the system automates product ranking and predicts price trends, enhancing both consumer convenience and decision-making efficiency. Key outputs include interactive graphs illustrating predicted price trajectories for various products.

## II. LITERATURE REVIEW

Research on web scraping and its applications has highlighted its role in collecting structured data from unstructured web content. Combining insights from various studies highlights the effectiveness of advanced web scraping methods [1], the role of scraping in structured data integration [2]

1) Aswad et al. (2023): Demonstrated the use of web scraping for product comparison websites, integrating recommendation algorithms to enhance user experience [1].
2) Srividhya et al. (2019): Emphasized data visualization post-web scraping to simplify consumer decision-making [2].
3) Bo Zhao (2017): Discussed advanced web scraping techniques involving tools like BeautifulSoup and Selenium [3].

These studies underline the growing relevance of web scraping in extracting actionable insights from large datasets. Combining this with machine learning amplifies its potential for innovative solutions in e-commerce.

## III. METHODOLOGY

### A. *Web Scraping*

The platform uses Python libraries such as BeautifulSoup and Selenium to scrape data from Amazon, eBay, Snapdeal, and Ajio. These tools are consistent with strategies discussed in research for efficient data extraction [1, 2]. The scraped data includes product titles, prices, ratings, and URLs. To ensure ethical scraping, the system adheres to site policies outlined in robots.txt files, aligning with ethical guidelines mentioned in recent studies [3].

### B. *Data Preprocessing*

Scraped data is cleaned to handle inconsistencies:

1) Prices are normalized by removing currency symbols and commas.
2) Missing ratings are replaced with default values.

### *C. Clustering and Ranking*

Clustering is performed using the K-Means algorithm to group products based on normalized price and rating. This approach is consistent with studies emphasizing machine learning's role in effective product sorting and ranking [1, 4]. Each product is assigned a predicted score based on its proximity to the centroid and user rating, ensuring relevance and accuracy in recommendations [3].

### *D. Price Prediction*

Linear regression models are employed to predict future price trends, a method validated by prior studies emphasizing the utility of predictive analytics for consumer insights [4]. This approach provides users with actionable predictions on cost changes over time, enhancing decision-making efficiency.

### *E. Flowchart of System Workflow*

The following Figure 1 flowchart illustrates the system workflow for product analysis. It depicts the stages from user input to detecting future price trends using linear regression.

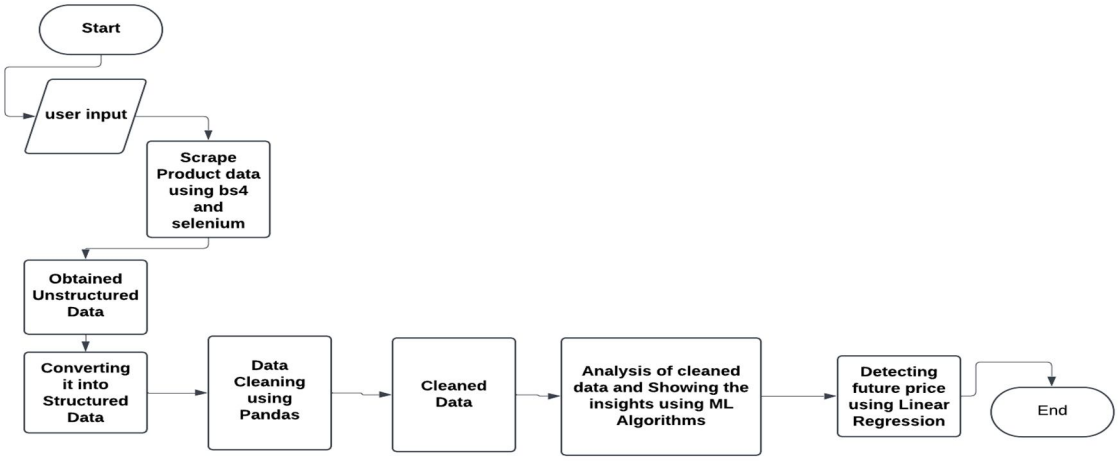



Fig. 1 Workflow of the Product Analysis System

## IV. RESULTS AND DISCUSSION

### *A. Data Extraction Performance*

The system successfully scraped entries across four platforms. Challenges included handling CAPTCHAs and dynamic JavaScript content.

### *B. Product Ranking*

Figure 2 shows K-Means clustering effectively segmented products into three clusters, highlighting affordability and quality differences. Consumers can easily identify products tailored to their preferences.


Fig. 2 Best three Products using K Means Clustering

### *C. Price Prediction Graph*

Figure 3 showcases the price prediction graph for Amazon products. The graph highlights:

*1)* Current price trends.
*2)* Predicted prices for the next 10 time steps.
*3)* Insights into seasonal price fluctuations or market trends.

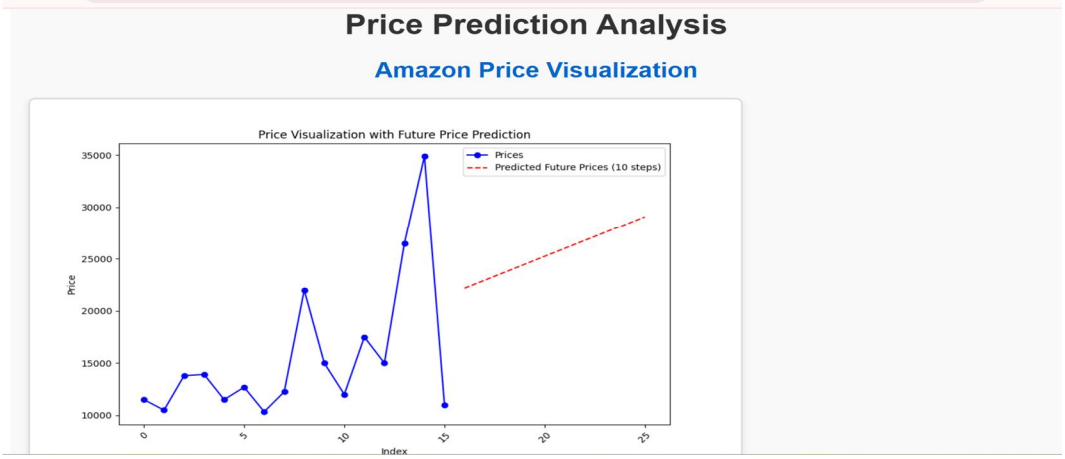



Fig. 3 The Price Predicted for the next 10 steps

### *D. Scraped Data Snapshot*

Figure 4 displays an example of scraped data from the system interface, showcasing extracted product details such as title, price, and ratings.

**Amazon Products**

| Title | Price | Rating | Link |
|---|---|---|---|
| Vivo T3 Lite 5G Smartphone (Vibrant Green, 6GB Ram 128GB Storage) | 11,497 | 3.3 | View Product |
| Vivo T3 Lite 5G Smartphone (Vibrant Green, 4GB Ram 128GB Storage) | 10,490 | 3.6 | View Product |
| Vivo T3x 5G (Celestial Green, 128 GB) (6 GB RAM) | 13,789 | 4.1 | View Product |
| Vivo T3x 5G (Crimson Bliss, 128 GB) (6 GB RAM) | 13,920 | 3.9 | View Product |
| Vivo T3 Lite 5G Smartphone (Majestic Black, 6GB Ram 128GB Storage) | 11,499 | 3.9 | View Product |
| Vivo T3X 5G (Crimson Bliss, 4GB Ram, 128GB Storage) | 12,690 | 4.0 | View Product |
| Vivo T3 Lite 5G Smartphone (Majestic Black, 4GB Ram 128GB Storage) | 10,338 | 3.3 | View Product |
| Vivo T3x 5G (Celestial Green, 128 GB) (4 GB RAM) | 12,246 | 4.1 | View Product |
| Vivo Y300 5G (Titanium Silver, 8GB RAM, 128GB Storage) with No Cost EMI/Additional Exchange Offers | 21,999 | 4.3 | View Product |
| vivo Y28s 5G (Vintage Red, 6GB RAM, 128GB Storage) with No Cost EMI/Additional Exchange Offers | 14,999 | 3.7 | View Product |
| vivo Y28e 5G (Vintage Red, 4GB RAM, 128GB Storage) with No Cost EMI/Additional Exchange Offers \| Without Charger | 11,999 | 3.7 | View Product |
| vivo Y58 5G (Sundarbans Green, 8GB RAM, 128GB Storage) with No Cost EMI/Additional Exchange Offers | 17,499 | 4.1 | View Product |

Fig. 4 The scraped data

### *E. Discussion*

While the system offers robust analytics, limitations include:

*1)* Data Completeness: Missing data from platforms due to restricted APIs.
*2)* Scalability: Scraping large datasets in real-time requires additional resources.

## V. CONCLUSION AND FUTURE SCOPE

This study demonstrates the efficacy of combining web scraping and machine learning to create an intelligent product comparison tool. The integration of clustering and price prediction provides users with a dynamic, data-driven shopping experience. Future enhancements could include:

*1)* Incorporating deep learning for more nuanced recommendations [4].
*2)* Expanding data sources to include user reviews and social media trends.
*3)* Developing mobile applications for broader accessibility.

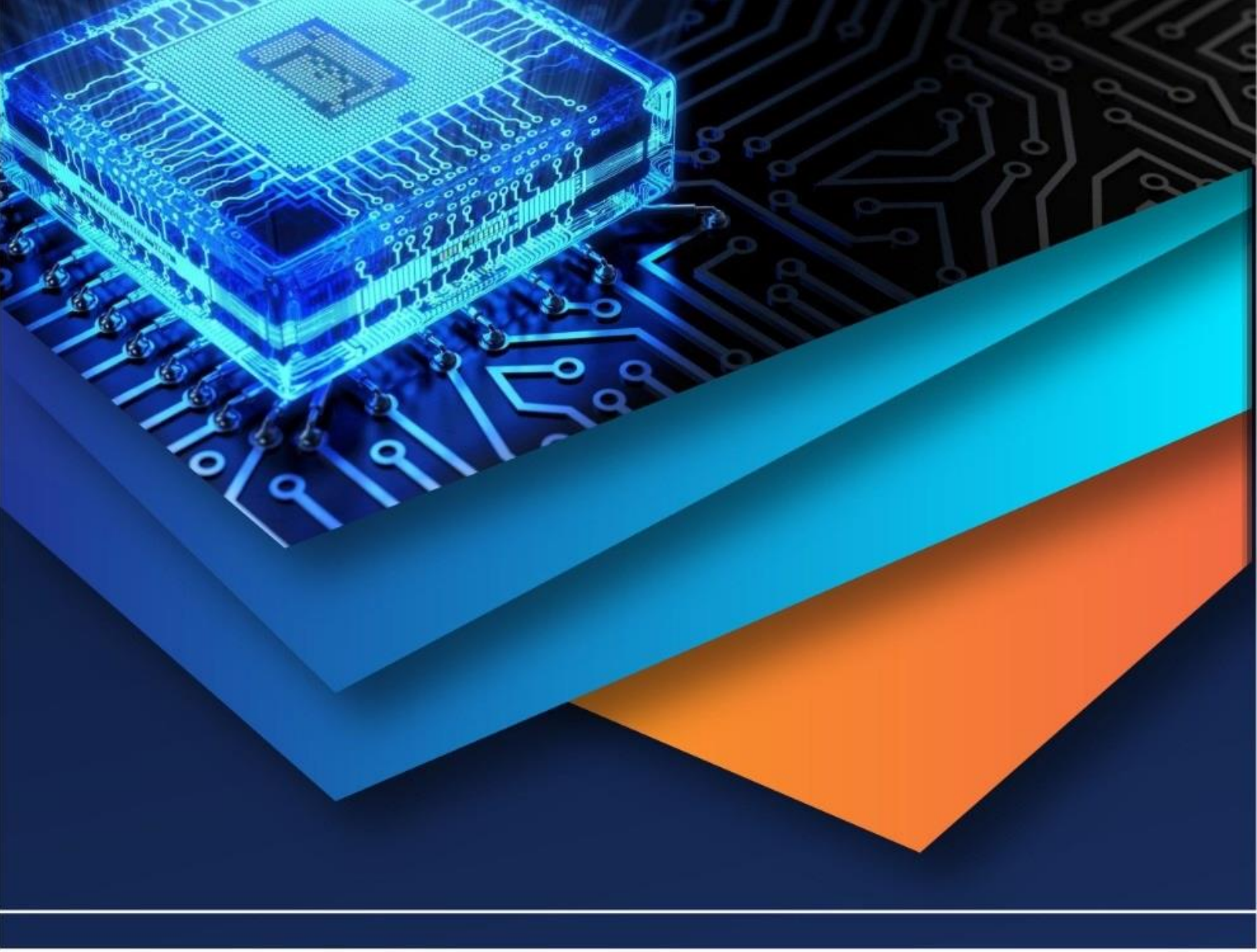
## VI. ACKNOWLEDGEMENT

We, the authors, express our heartfelt gratitude to Jawaharlal Nehru New College of Engineering and the Department of Computer Science and Engineering for providing the resources and support required to carry out this work.

## REFERENCES

[1] Aswad Shaikh, Aniket Sonmali, and Soham Wakade, "Product Comparison Website using Web Scraping and Machine Learning," Student at Department of Information Technology, Atharva College of Engineering, Mumbai, Maharashtra, India.
[2] V. Srividhya and P. Megala, "Scraping and Visualization of Product Data from E-commerce Websites," Dept of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamilnadu, India.
[3] B. Zhao, "Web Scraping," College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, OR, USA.
[4] C. Lotfi, S. Srinivasan, M. Ertz, and I. Latrous, "Web Scraping Techniques and Applications: A Literature Review," Labo NFC, University of Quebec at Chicoutimi, 555 Boulevard de l'Université, Saguenay (QC), Canada.
[5] M. Fraňo, "Web Scraping as a Data Source for Machine Learning Models and the Importance of Preprocessing Web Scraped Data," University of Twente, The Netherlands.
[6] S. C. M. de S. Sirisuriya, "Importance of Web Scraping as a Data Source for Machine Learning Algorithms - Review," Department of Computer Science, Faculty of Computing, General Sir John Kotelawala Defence University, Sri Lanka.
[7] R. Praba, G. Darshan, K. T. Roshanraj, and B. Surya Prakash, "Study On Machine Learning Algorithms," Department of Information Technology, Dr. N.G.P. Arts and Science College, Coimbatore, Tamil Nadu, India.
[8] T. S. Singh and H. Kaur, "Review Paper on Django Web Development," Department of Computer Application, Rayat Bahra University.
[9] Data Visualization workshop, Tim Grobmann and Mario Dobler, Packt Publishing, ISBN 9781800568112
[10] https://docs.python.org/3/
[11] https://pypi.org/project/beautifulsoup4/
[12] https://pypi.org/project/requests/
[13] https://www.djangoproject.com/
[14] https://www.selenium.dev/documentation/
[15] https://docs.djangoproject.com/en/stable/
[16] Adrian Holovaty, Jacob Kaplan Moss, The Definitive Guide to Django: Web Development Done Right, Second Edition, Springer-Verlag Berlin and Heidelberg GmbH & Co. KG Publishers, 2009
[17] S. Sridhar, M Vijayalakshmi "Machine Learning",Oxford,2021.
[18] https://scikit-learn.org/stable/documentation.html



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY