IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 2, February 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 2, February 2024 ||

| DOI:10.15680/IJIRSET.2024.1302018 |

Analysis of Google Search Using Machine Learning Techniques

Shaili Singh[1], Vansh Goel[2], Mr. Gautam Kumar[3]

U.G. Students, School of Computer Science and Engineering, Galgotias University, Greater Noida, India[1,2]

A.P., School of Computer Science and Engineering, Galgotias University, Greater Noida, India[3]

ABSTRACT: This research paper presents an integrated methodology for analysing Google Trends data, leveraging a combination of machine learning techniques and time series forecasting. The study aims to provide a comprehensive understanding of user interest trends over time, offering insights that can inform decision-making processes. The approach involves the application of Linear Regression, Random Forest Regression, and Support Vector Regression (SVR) for machine learning modelling, alongside the Autoregressive Integrated Moving Average (ARIMA) model for time series forecasting. The project's user interface, implemented using Tkinter, enhances accessibility, allowing users to seamlessly interact with the analysis. Robust error handling mechanisms contribute to a positive user experience. The results demonstrate the effectiveness of each model in predicting user interest trends for Artificial Intelligence, Machine Learning, and Data Science. Linear Regression showcases its prowess in simplicity, Random Forest Regression captures intricate relationships, SVR effectively handles non-linear patterns, and ARIMA excels in time series forecasting. The paper concludes with insights drawn from the comparative analysis and outlines future directions for research and development.

KEYWORDS: Google Trends, Machine Learning, Linear Regression, Random Forest Regression, SVR, ARIMA.

I. INTRODUCTION

In the digital age, understanding user behaviour and interests is paramount for various applications, ranging from strategic business planning to content creation. Google Trends data provides a valuable source of information, reflecting the dynamic nature of user interests over time [1].This research aims to harness the potential of Google Trends data through an integrated analysis, blending machine learning techniques with time series forecasting.

By employing machine learning algorithms such as Linear Regression, Random Forest Regression, and Support Vector Regression, we seek to model and predict user interest trends for specified search terms. The comparative analysis of these models provides insights into their respective accuracies and performance [2].

In addition, the project incorporates time series forecasting using the Autoregressive Integrated Moving Average (ARIMA) model. This approach allows us to not only understand historical trends but also predict future values of user interest, contributing to a comprehensive understanding of the dynamic nature of online user behaviour.

The integration of machine learning and time series forecasting in this project provides a versatile tool for analysts and researchers, offering a full view of Google Trends data [3]. The user-friendly interface enhances accessibility, making it a valuable resource for decision-makers navigating the ever-changing digital landscape [4].

II. LITERATURE SURVEY

In the era of the rapidly expanding digital landscape, the utilization of search engines has become increasingly integral to online activities. Notably, popular search engines like Google, YouTube, and Bing have gained prominence owing to their sophisticated search algorithms, ensuring the delivery of highly relevant results to users [5]. The effectiveness of these platforms lies in their adept searching techniques, analysing diverse elements within a webpage's source code, including title, headings, descriptions, time of last update, mobile design, and structured data for rich snippets [6].

Studies indicate that user engagement with websites is a multi-visit process, often involving multiple interactions before a conclusive action, such as making a purchase decision, occurs [7]. Central to this engagement is the pivotal

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 2, February 2024 ||

| DOI:10.15680/IJIRSET.2024.1302018 |

role of a website's content, which not only attracts users but also sustains their interest over repeated visits. This underscores the importance of crafting websites with targeted content that aligns with user interests and ranks favourably in search engine results [8].

The dynamics of search engines, such as Google and Yahoo, involve the use of spiders for crawling and indexing web pages. This process encompasses both new and updated pages, facilitating an indexing mechanism that enhances user search experiences [9]. During indexing, web content undergoes parsing, extracting valuable information from elements like title, description, URL, and hyperlinks [10]. This extracted information contributes to the generation of relevant indexing data, optimizing the efficiency of search engine queries.

In the context of the presented literature survey, the focus shifts to the exploration of Google Trends data, integrating machine learning techniques and time series forecasting. This research aims to provide a comprehensive understanding of user interest trends over time, leveraging methods such as Linear Regression, Random Forest Regression, Support Vector Regression (SVR), and the Autoregressive Integrated Moving Average (ARIMA) model. The user-friendly interface implemented through Tkinter enhances accessibility, allowing users to interact seamlessly with the analysis. The comparative analysis of machine learning models and time series forecasting using ARIMA provides insights into the efficacy of each method, offering a versatile tool for researchers and analysts navigating the dynamic landscape of online user behaviour.

This literature survey sets the stage for a comprehensive exploration of the research project, emphasizing the evolution of search engines, user engagement with web content, and the integration of advanced analytics techniques to decipher patterns in Google Trends data.

III. METHODOLOGY

The methodology involves the acquisition of Google Trends data, pre-processing steps such as date conversion and feature engineering, and the application of machine learning models. The initial phase involves the utilization of the `pytrends` library to fetch Google Trends data, allowing for an unbounded exploration of user search behaviour. Machine learning techniques form the cornerstone of the analytical approach [11]. The implementation encompasses rigorous data pre-processing steps to ensure the reliability and quality of the Google Trends data. The comparative analysis of Linear Regression, Random Forest Regression, and SVR provides insights into their respective strengths.

Additionally, time series forecasting using the ARIMA model contributes to the project's holistic approach. The user interface implemented using Tkinter facilitates user interaction, and robust error handling mechanisms ensure a smooth user experience.

1. Acquisition of Google Trends Data: The initial phase involves the acquisition of Google Trends data, a pivotal step in understanding user search behaviour [10]. The `pytrends` library is utilized to fetch historical interest data for specified search terms, providing a rich dataset for analysis.

2. Pre-processing Steps: Rigorous pre-processing is undertaken to ensure the reliability and quality of the Google Trends data. This includes essential steps such as date conversion and feature engineering [12]. The 'date' column is transformed into datetime format, and additional temporal features, such as month and day, are extracted to enhance the dataset's analytical depth.

3. Linear Regression Modelling: Linear Regression is employed as a foundational machine learning model to predict user interest trends. This model leverages the relationship between the temporal features (month, day) and the target variable (search term interest). The training of the Linear Regression model is executed on the pre-processed data, and subsequent predictions are evaluated to gauge its accuracy using metrics such as Mean Squared Error (MSE) [13] [14].

4. Random Forest Regression Modelling: In parallel with Linear Regression, Random Forest Regression is implemented to capture more complex relationships within the data [13]. This ensemble model constructs multiple decision trees and aggregates their predictions. The training process involves exposing the model to the pre-processed data, and the accuracy is assessed through evaluation metrics like MSE [14].

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 2, February 2024 ||

| DOI:10.15680/IJIRSET.2024.1302018 |

5. Support Vector Regression (SVR): To address non-linear patterns in the data, Support Vector Regression (SVR) is applied. SVR is a robust regression technique that excels in capturing intricate relationships [13] [15]. The model is trained using the pre-processed data, and its effectiveness is evaluated based on predictive accuracy measured by MSE [16].

6. Time Series Forecasting with ARIMA: The time series forecasting component employs the Autoregressive Integrated Moving Average (ARIMA) model. ARIMA is particularly suited for capturing temporal dependencies and trends in time series data. The model is tailored for each search term based on historical data, and its accuracy is evaluated through forecasted values and MSE.

7. User Interface Implementation using Tkinter: The project's user interface is implemented using Tkinter, a Python library for GUI development. This facilitates seamless user interaction, allowing users to input search terms and date ranges effortlessly. The Tkinter interface enhances the accessibility of the analysis, catering to users with varying levels of technical expertise.

8. Robust Error Handling Mechanisms: Anticipating challenges such as data variability and API limitations, robust error handling mechanisms are integrated. These mechanisms ensure that discrepancies in user inputs or potential issues with API interactions are gracefully handled, contributing to an overall positive user experience [16].

By systematically incorporating these subtopics into the methodology, the research establishes a robust framework for analysing Google Trends data, encompassing machine learning modelling, time series forecasting, user interface design, and error handling. Each subtopic contributes uniquely to the project's objectives, offering a holistic and insightful exploration of user interest trends over time.

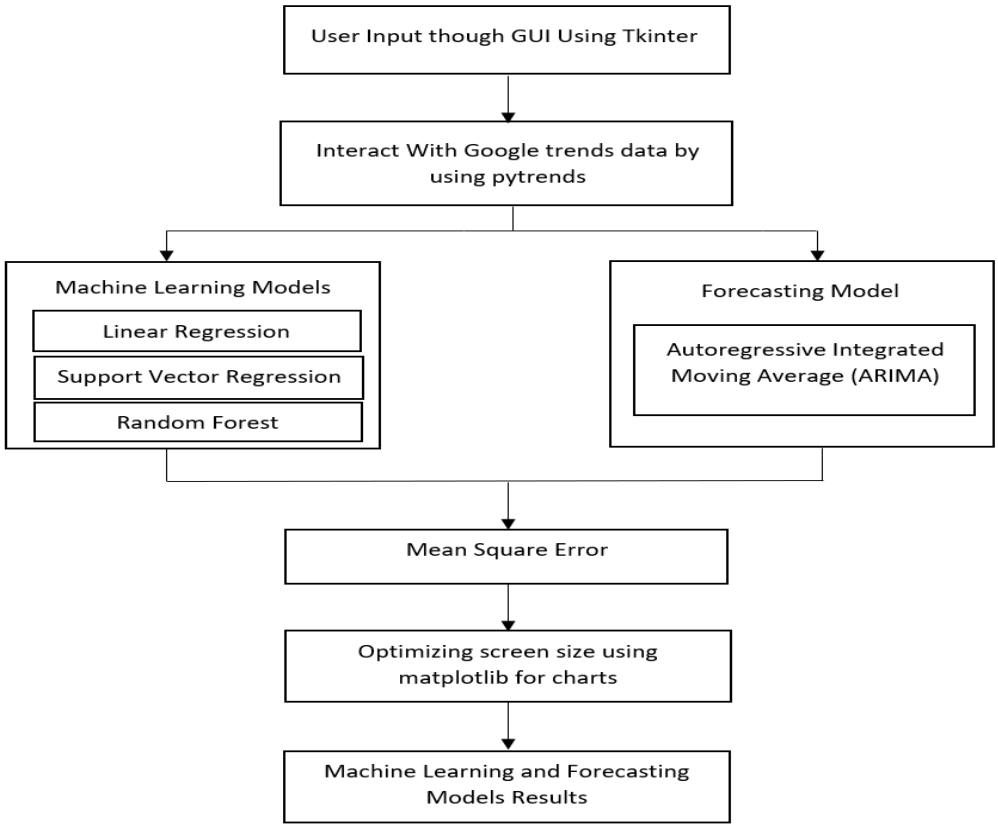


Figure 1: DFD for model working

Anticipating challenges such as data variability and rate limitations imposed by Google, the research is poised to address these hurdles through meticulous pre-processing and efficient handling of API interactions. The absence of

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

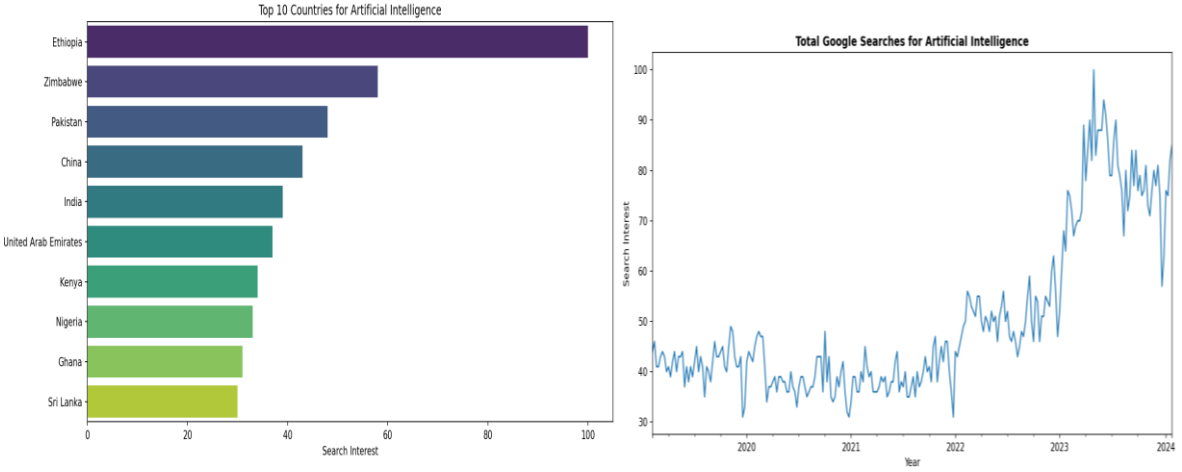
|| Volume 13, Issue 2, February 2024 ||

| DOI:10.15680/IJIRSET.2024.1302018 |

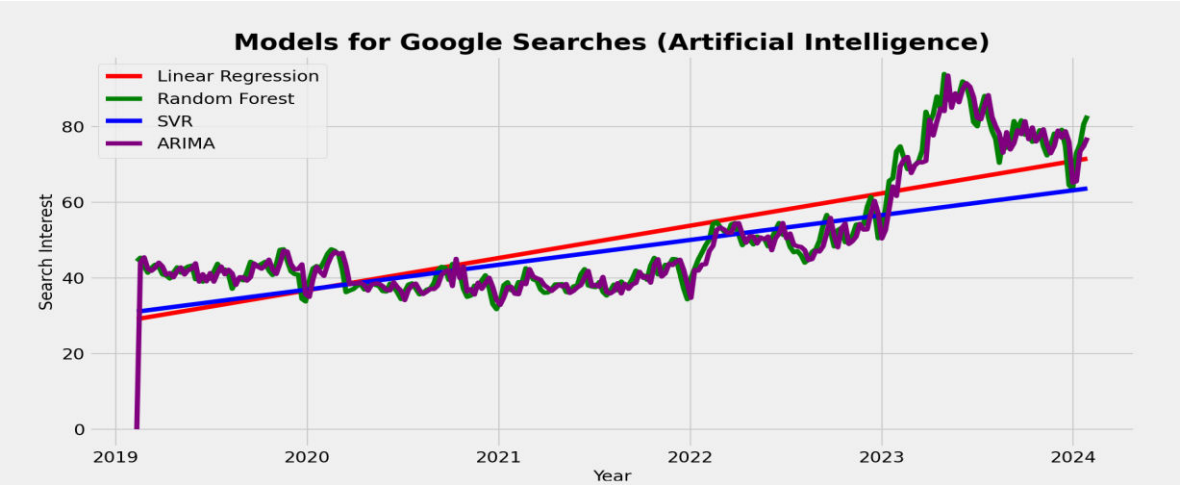
predefined search terms and time periods, while presenting challenges, also opens avenues for a nuanced exploration of user behaviour on a grand scale.

IV. RESULTS

The results showcase the top 10 countries that most searches the keyword and total google searches frequency graph and effectiveness of each machine learning model and the ARIMA model. Linear Regression excels in scenarios with linear trends, while Random Forest Regression captures complex relationships. SVR proves effective in handling non-linear patterns. The ARIMA model accurately forecasts future values of user interest trends, enhancing the project's predictive capabilities.



(a) (b)



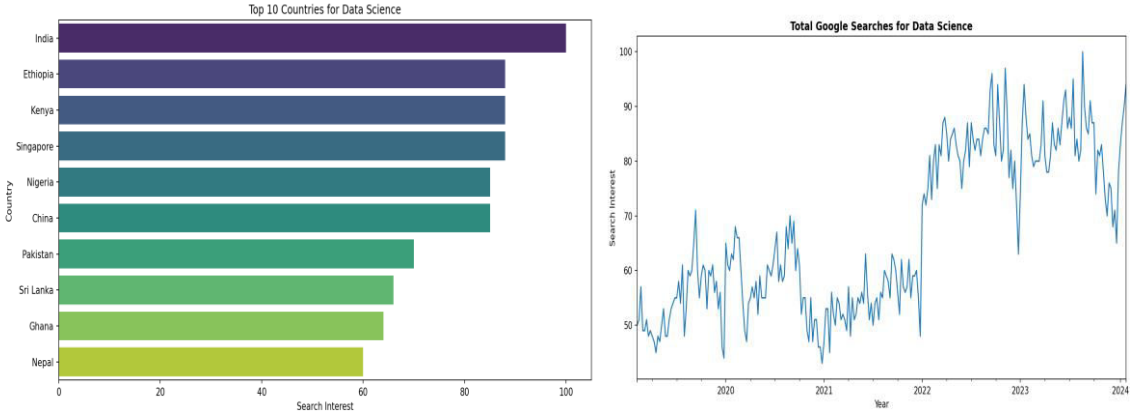
(c)

Figure 2: Artificial Intelligence (a) top 10 countries (b) Total Google Search (c) Machine Learning Models

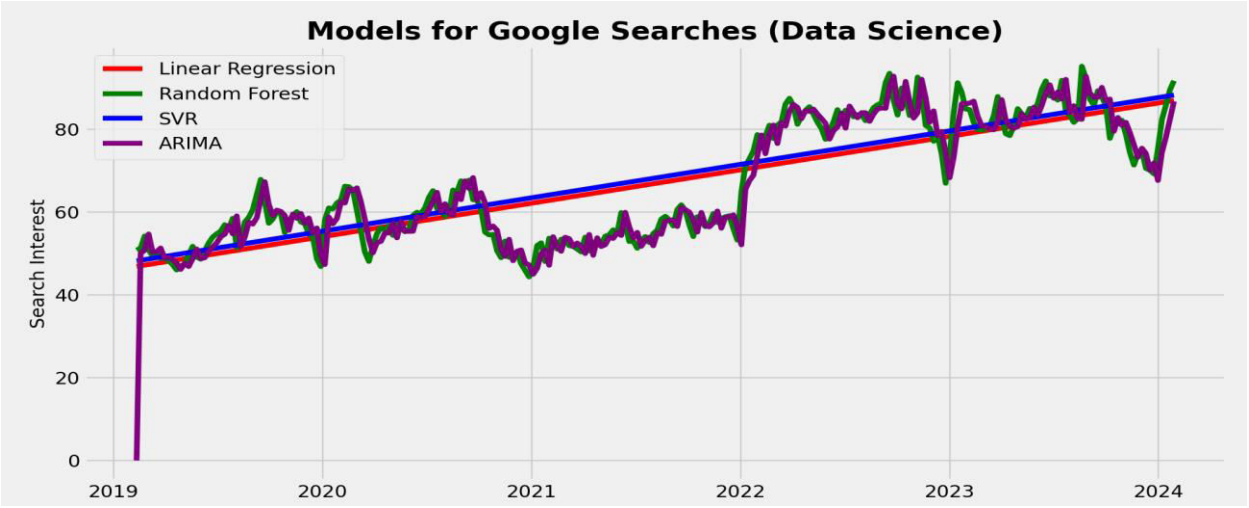
|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 2, February 2024 ||

| DOI:10.15680/IJIRSET.2024.1302018 |

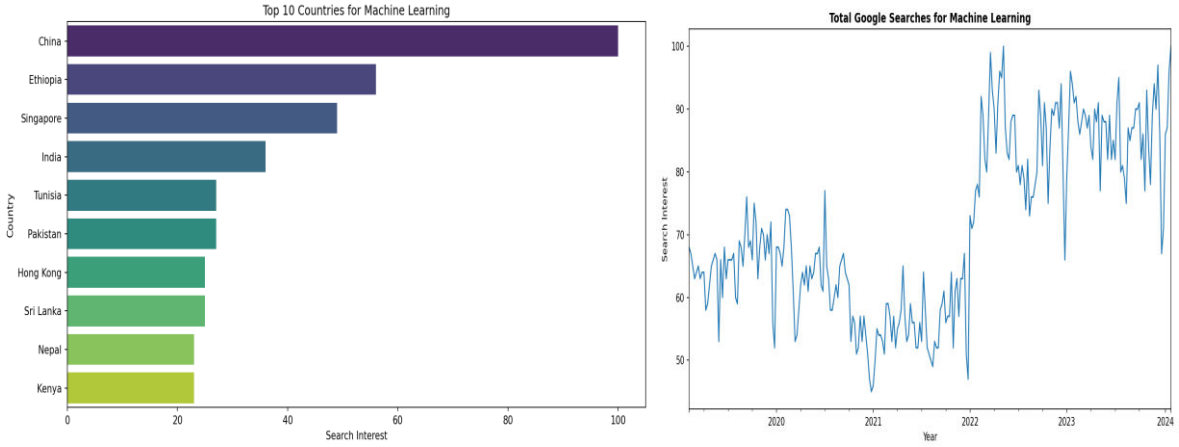


(a) (b)



(c)

Figure 3: Data Science (a) top 10 countries (b) Total Google Search (c) Machine Learning Models

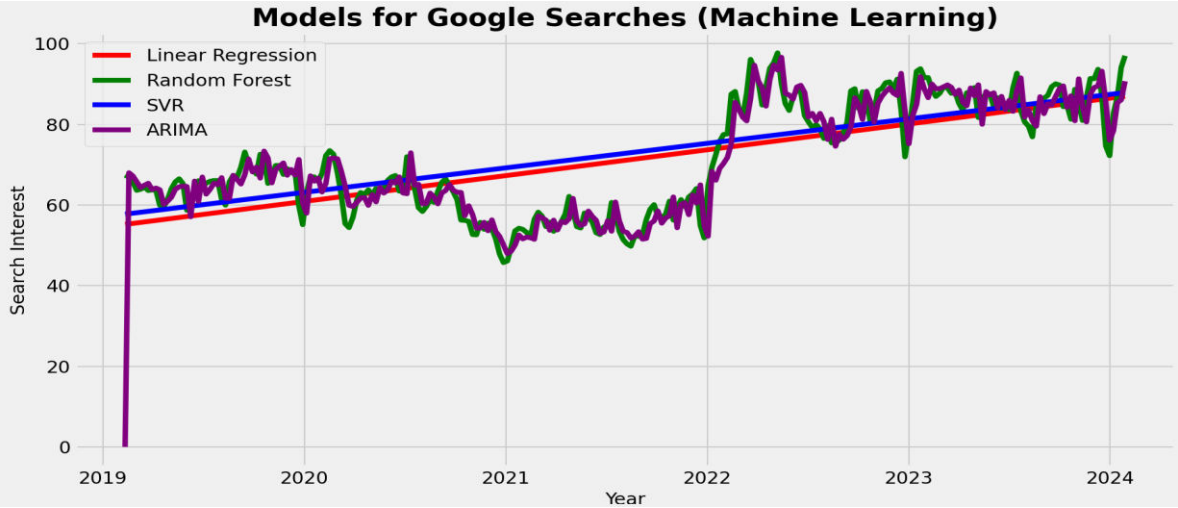


(a) (b)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 2, February 2024 ||

| DOI:10.15680/IJIRSET.2024.1302018 |



(c)

Figure 2: Artificial Intelligence (a) top 10 countries (b) Total Google Search (c) Machine Learning Models

The ARIMA model, represented by the lines, showcases its proficiency in forecasting future user interest trends. The figure juxtaposes the forecasted values with the actual observed values, providing a comprehensive view of the model's predictive capabilities.

4.5 Mean Squared Error (MSE) Comparison: Table 1 presents the Mean Squared Error (MSE) values for each machine learning model across the selected search terms.

Model	Artificial Intelligence	Machine learning	Data Science
Linear Regression	41.54177420101608	89.18120439733764	241.58738610828732
Random Forest	58.47362820160114	133.43052035038198	341.98482255571525
SVR	20.57030435485694	71.0290020198607	184.67029162647253
ARIMA	80.83224842354528	273.12838825589466	371.7255999524117

Table 1:Mean Square Error Comparison for Search Terms

The MSE values provide insights into the relative performance of each model, with lower values indicating better predictive accuracy.

V. CONCLUSION AND FUTURE WORK

In conclusion, this project successfully integrated machine learning and time series forecasting to analyse Google Trends data, shedding light on user interest trends over time. The combination of Linear Regression, Random Forest Regression, Support Vector Regression (SVR), and the ARIMA model allowed for a comprehensive understanding of both linear and non-linear patterns in the data.

The comparative analysis revealed the strengths of each machine learning model, with Linear Regression excelling in simplicity, Random Forest Regression capturing complex relationships, and SVR handling non-linear patterns effectively. The ARIMA model showcased its capability in time series forecasting, providing accurate predictions of future user interest trends.

The user interface implemented using Tkinter enhances the project's usability, allowing users to interact with the analysis seamlessly. Robust error handling mechanisms contribute to a positive user experience by providing informative messages in case of input discrepancies.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 2, February 2024 ||

| DOI:10.15680/IJIRSET.2024.1302018 |

Future work could explore ensemble modelling, hyperparameter tuning for machine learning models, dynamic model updating, integration of additional features, and enhancements to the user interface. Deploying the project as a service could further extend its accessibility and usability.

REFERENCES

[1] M. D. Y. Pawade, "Analyzing the Impact of Search Engine Optimization Techniques on Web Development Using Experiential and Collaborative Learning Techniques," Modern Education and Computer Science, pp. 1-10, 2021.
[2] J. J. J. Sojung An, "A heuristic approach on metadata recommendation for search engine optimization," in ICIDB , Seoul, Korea, January 2019.
[3] Z. S. Y. Y. Mengjie Si, "The Analysis of Google's Monopoly in The Search Engine Industry," ICFIED , vol. 211, pp. 2168-2173, 2022.
[4] I. J. a. P. P. Rogers Kaliisa, "A checklist to guide the planning, designing, implementation, and evaluation of learning analytics dashboards," International Journal of Educational Technology in Higher Education, 03, May 2023.
[5] H. N. Murtaza Ahmed Hingoro, "A Comparative Analysis of Search Engine Ranking Algorithms," IJATCSE, vol. 10, no. 2, pp. 1247-1252, April 2021.
[6] A. S. A. P. RUTECKA, "Direct Answers in Google Search Results," the Ministry of Science and Higher Education of the Republic of Poland, vol. 8, pp. 103642-103654, 2020.
[7] C. Z. a. G. Wu, "Research and Analysis of Search Engine Optimization Factors Based on RE," ICMINS, pp. 226-228, 2011.
[8] M. P. Evans, "Analysing Google rankings through search engine optimization data," INTR, vol. 17, no. 01, pp. 21-37, 2007.
[9] R. S. S. K. A. K. G. Dushyant Sharma, "A BRIEF REVIEW ON SEARCH ENGINE OPTIMIZATION," in International Conference on Cloud Computing, Data Science & Engineering, 2019.
[10] B. O. a. C. T. Lopes, "The Evolution of Web Search User Interfaces - An Archaeological Analysis of Google SERP," in CHIIR, Austin, TX, USA, 2023.
[11] J. P. D. P. Akshita Patil, "Comparative Study Of Google SEO: Panda, Penguin and Hummingbird," in I2CT, Pune, India, April, 2021.
[12] D. Mrąbek, "The Google Search Engine: A Blended-Learning Tool for Student Empowerment," ISET, pp. 224-229, 2019.
[13] J. D. a. D. M. Goran Matoševic, "Using Machine Learning for Web Page Classification in SEO," MDPI, vol. 13, no. 9, pp. 1-20, 2021.
[14] S. Q. a. L. Chiang, "Advances and opportunities in machine learning for process data analytics," ELSEVIER, pp. 465-473, 2019.
[15] V. R. K. a. V. S. Kotapati Narayana Loukika, "Analysis of Land Use and Land Cover Using ML Algorithms on Google Earth Engine," MDPI, vol. 13, p. 13758, 2021.
[16] G. M. M. N. S. A. P. V. M. Ch.Venkata Ramana, "Building Search Engine Using Machine Learning Technique," IJIRT, vol. 8, no. 4, pp. 172-176, 2021.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details