ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

NEXT-GENERATION ENGAGEMENT FORECASTING FOR INSTAGRAM USING EXPLAINABLE MACHINE LEARNING MODELS

K. JAYA SRI[1], POTHIPOGU YESUBABU[2]

[1] Assistant Professor, Department of MCA, Chirala Engineering College, Chirala, Andhra Pradesh, India-523157.

Email: srij4620@gmail.com

[2] MCA Student, Department of MCA, Chirala Engineering College, Chirala, Andhra Pradesh, India-523157.

Email: yesubabupothipogu4@gmail.com

ABSTRACT

The increasing popularity of social media platforms, especially Instagram, has reshaped online communication and digital marketing, where user engagement now serves as a key indicator of content performance. Instagram has evolved from a simple photo-sharing application into a highly interactive platform where businesses, influencers, and organizations depend on engagement metrics such as likes, comments, shares, and reach to assess content performance. As the volume of content and user interactions continues to grow, analysing and predicting engagement rates has become increasingly challenging due to the complex relationships among user behaviour, content characteristics, and platform dynamics. The primary problem addressed in this research is the difficulty of accurately predicting Instagram engagement rates using traditional analytical approaches. Conventional methods rely on platform-generated statistics, manual analysis, and static reports, which provide descriptive insights but lack predictive capabilities. These approaches are often limited by poor scalability, inability to identify hidden patterns, delayed decision-making, and dependence on subjective interpretation. To overcome these limitations, the proposed system implements a machine learning-based framework for predictive modelling of Instagram engagement rates. The framework includes data preprocessing, exploratory data analysis, feature scaling, model training, and performance evaluation. Several regression algorithms, including Linear Regression (LR), K-Nearest Neighbours Regressor (KNNR), Random Forest Regressor (RFR), and Decision Tree Regressor (DTR), are developed and compared using evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), and R^2 Score. Experimental results indicate that tree-based models outperform other techniques, with the Decision Tree Regressor achieving the highest prediction accuracy. The proposed system provides an effective, scalable, and data-driven solution for engagement prediction, supporting improved content strategies and informed decision-making in social media analytics.

Keywords: Machine Learning, Regressor Models, Decision Tree Regressor, Engagement Forecasting, Graphical User Interface.

1. INTRODUCTION

Post user engagement on social media platforms refers to the interactions a post receives from its audience, serving as a key metric for measuring content effectiveness and audience interest. Engagement encompasses a variety of elements, including likes, comments, shares, saves, and click-throughs, each reflecting a different type of user interaction [1]. These elements provide insights into user preferences and behaviour, allowing content creators, brands, and marketers to evaluate the success of their strategies [2]. Understanding post engagement is essential for optimizing content to foster stronger connections with the target audience and achieve specific goals, such as increasing visibility, driving traffic, or enhancing brand loyalty [3]. This paper examines post engagement as a multifaceted metric and explores predictive models to understand better its relationship with account follower counts [4]. Presently, the social network market grows both in number of operators and in number of posts, exponentially. With millions of monthly active users, both on mobile devices and web browsers, Instagram represents a leading platform in this market [5].

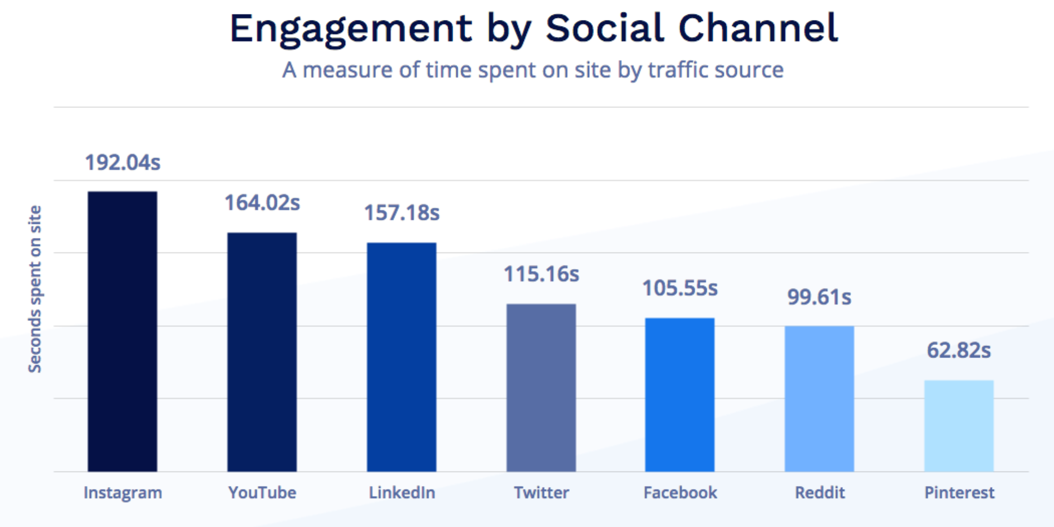


Fig. 1: Instagram engagement rate

Launched in 2010, it has gradually gained a leading role among photo-sharing platforms, introducing several innovative features over time, including not exhaustively filters, stories, and an internal messaging system as shown in the fig 1 [6]. Recently, these features have attracted not only ordinary users and photography enthusiasts, but also companies, organizations and global brands, thanks to the possibility that Instagram has offered to explore new business models and marketing strategies [7]. In such a context, approaches aimed to predict the popularity of social content are gaining increasing attention, not only from a commercial and industrial standpoint, but also from a scientific perspective [8]. Indeed, thanks to the advancement of knowledge in the fields of data analysis and artificial intelligence, together with the development of new techniques based on Machine learning, Data mining, Big Data, etc., it is now possible to provide advanced tools for companies and individuals and promote the consolidation of these businesses in the social market [9].

2. LITERATURE SURVEY

Padiya, et al. [10] represented a comprehensive approach toward classification of Instagram users and the prediction of engagement rates by state-of-the-art machine learning techniques. High variance in engagement rates from fake and inactive users mostly acts as a barrier to the learning process of the model. In this regard, they proposed a new approach to reduce these variations to enhance the performance of the model.

Son, et al. [11] analysed the entire three major online travel agencies Instagram posts (n = 6,083) to investigate which features contribute more to predicting the user engagement. Among 109 textual, visual, and social media post specific features that they initially extracted, they find the important features using the XGBoost algorithm and estimate the effects of each feature on user engagement (i.e. number of likes) using Negative Binomial regression.

Yew, et al. [12] aimed of gaining a wider range of knowledge of how to determine the efficiency of influencers. research has been done about related case studies and existing algorithms to get a better understanding of how to use available metrics such as likes, comments and followers, to calculate the engagement rate of influencers in the most accurate way. Brand owners can then select the most efficient influencer to advertise their product.

AlAnezi, et al. [13] proposed a more sophisticated engagement rate measure based on a careful analysis of potential customers reaction to an influencer advertisement post. The new measure works only on advertisements posts. Also, it considers the polarity of the comments on these posts, not only their count. To efficiently compute the new engagement rate measure over large size of data, they proposed machine learning (ML) approaches to generate necessarily information to compute the new engagement rate measure.

Marimuthu, et al. [14] analysed revealed significant correlations between various engagement metrics, such as a linear relationship between likes and reach, and weaker yet notable impacts of comments and shares. Visualizations provided actionable insights into content strategies, including the effectiveness of specific hashtags and caption styles. The Passive Aggressive Regressor demonstrated adaptability and accuracy, supported by evaluation metrics like R-squared and Mean Absolute Error. The model successfully predicted engagement metrics and offered strategic recommendations for content optimization.

Mazloom, et al. [15] compared several regression techniques to predict the Engagement Rate (ER) of posts using a global dataset. The prediction model, coupled with the results of the popularity trend analysis, will have more utility for a larger audience compared to existing studies. The features were extracted from hashtags, image analysis, and user history.

Ramya, et al. [16] investigated user engagement on Instagram using Instagram Reach Analysis Data, focusing on key metrics such as likes, comments, shares, and views to understand content performance and user interaction patterns. they employed Python-based data analysis, utilizing Pandas for data extraction and visualization tools, including histograms and scatter plots, to interpret the data. Additionally, they used NetworkX for network analysis to explore the impact of different sources of impressions such as home feeds, hashtags, and the explore page on total reach. Through correlation analysis and graph-based visualizations, they uncovered intricate relationships among various engagement metrics.

Bigdeli, et al. [17] conducted in a semi-structured manner, allowing respondents to contribute to this process. The findings of the study reveal a total of 54 distinct customer behaviour patterns. The study demonstrated that product-related factors, content characteristics, technology usage, social influences, and situational context are the primary drivers of consumer behaviour on the Meta platform. It is noteworthy that the research findings indicate the considerable impact of product attributes, including pricing, quality, and product category, on consumer behaviour on Meta platform.

Oliveira, et al. [18] revealed that both kinds of features are relevant to predicting the success of a post, and the results differ according to the size of the follower base of the influencers. they hope that these and other results of our study provided valuable information about influence dynamics and strategies for content personalization on online social media platforms.

Dadgar, et al. [19] introduced a pipeline that ranks faces based on the number of times they appear on Instagram accounts using face clustering. Moreover, they conducted our analysis and experiments on a dataset of Instagram posts that they collected hashtags related to brands and celebrities. Unlike the existing works, they analysed these contents from both content and user perspectives and show a significant difference between data. Considering our results, they showed that our multimodal model outperforms other models and the effectiveness of object detection in detecting mismatched information.

Fan, et al. [20] discussed the benefits of multi-modal data integration, where combining different types of data leads to more robust models capable of capturing the complexities of user engagement and content virality. Additionally, the integration of reinforcement learning and real-time prediction systems is highlighted as an emerging direction, enabling models to adapt and improve based on real-time feedback. As deep learning continues to evolve, it offers significant potential for more accurate, scalable, and adaptive models that can respond to changing trends and enhance content dissemination strategies across social media platforms.

3. PROPOSED METHODOLOGY

The methodology followed in this research adopts a clear and systematic data-driven approach to understand and predict Instagram engagement rate using user and content-related features has shown in fig 2. The research starts with collecting Instagram analytics data that includes both numerical values and categorical attributes associated with posts and user interactions. This data is carefully cleaned and processed to maintain accuracy, consistency, and reliability before analysis. EDA is then performed to observe patterns, trends, and relationships within the dataset. Multiple machine learning regression models are trained and evaluated to compare their ability to predict engagement effectively. Standard evaluation metrics and visual comparisons help in understanding model performance and reliability. The trained models and prediction results are stored for reuse, enabling consistent analysis and practical applicability in real-world scenarios.

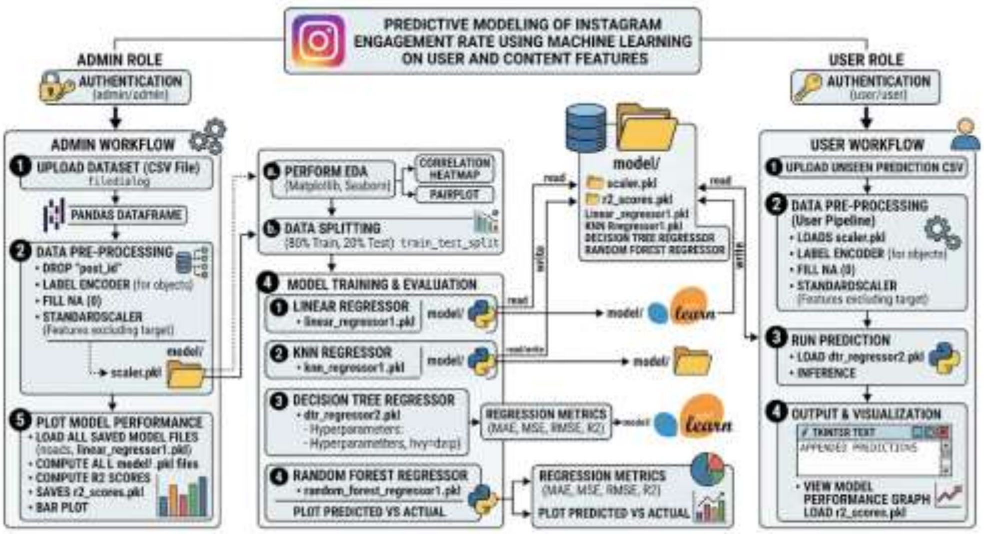


Fig. 2: System architecture of Instagram engagement rate

User Interface (Data Analyst / User)

- The user interacts with the system by providing Instagram datasets in CSV format.
- They can view exploratory data analysis outputs, correlation plots, and model evaluation results.
- The user initiates model training and test data prediction processes.
- All interactions are performed in a controlled analytical environment such as a Python notebook or application interface.

Data Source (Instagram Dataset – CSV Input)

- The dataset serves as the primary input to the system.
- It contains user attributes, content-related features, and engagement metrics.
- Both numerical and categorical data are included for comprehensive analysis.
- The data is forwarded to the preprocessing module for further handling.

Data Preprocessing & Feature Engineering

- Raw data undergoes cleaning to handle missing values and duplicate records.
- Categorical attributes are transformed using label encoding.
- Numerical features are standardized using scaling techniques.
- The final processed feature set is prepared for machine learning model input.

Exploratory Data Analysis Module

- Statistical summaries and visualizations are generated to understand data distribution.
- Correlation heatmaps reveal relationships between engagement rate and other features.
- Pair plots and box plots help identify patterns and outliers.
- These insights guide better interpretation of model behaviour.

Machine Learning Models

- The processed dataset is passed to multiple regression models for training:
    - KNNR
    - LR
    - RFR
    - DTR
- Each model learns engagement patterns independently.
- Predictions are generated for unseen test data.

Model Evaluation & Comparison

- Model predictions are evaluated using MAE, MSE, RMSE, and R^2 metrics.
- Performance results are stored for comparison across models.
- Graphical comparisons help identify the most reliable regression model.

- The best-performing model is selected for final prediction.

Model Storage & Prediction Output

- Trained models are saved using a lightweight storage mechanism.
- The selected model is used to predict engagement rate for new test data.
- Predicted values are appended to the dataset for interpretation.
- Results support data-driven insights into Instagram engagement behaviour.

Decision Tree Regressor (DTR)

DTR predicts engagement rate by learning a sequence of decision rules from data, as illustrated in fig. 3. The model splits the dataset based on feature conditions that reduce prediction error. Each split leads closer to a final prediction stored at leaf nodes. This structure makes the model easy to interpret. However, deep trees may overfit the data. Controlled depth improves generalization.

Selecting the Best Feature to Split: The algorithm starts by examining all input features to find the one that best separates the data. It evaluates possible split points and measures how much each split reduces prediction error. Commonly, variance reduction or mean squared error is used for this decision.

Splitting the Dataset: Once the best feature and split value are chosen, the dataset is divided into two or more subsets. Each subset contains data points that satisfy the split condition. This step creates branches in the tree structure.

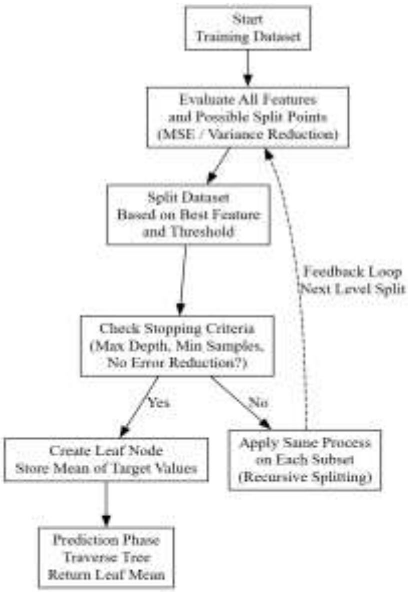


Fig. 3: Internal workflow of DTR

Repeating the Process Recursively: The same splitting logic is applied again to each subset created. The tree grows deeper if splitting improves prediction accuracy. This recursive process helps the model capture complex patterns in the data.

Stopping the Tree Growth: Tree growth stops when a stopping condition is met, such as reaching a maximum depth or having too few samples in a node. This prevents the model from becoming overly complex. Proper stopping helps reduce overfitting.

Making Predictions at Leaf Nodes: At the final nodes (leaf nodes), no further splitting occurs. The model stores the average of the target values in each leaf. When a new input arrives, it follows the tree rules and outputs the stored average as the prediction.

Key Points of DTR

- Easy to interpret
- Handles non-linear data
- No need for scaling
- Supports mixed data types
- Visual decision rules
- Fast prediction

Advantages of DTR

- Easy to understand and interpret, as predictions are made using clear decision rules.
- Handles non-linear relationships effectively without requiring complex transformations.
- Does not require feature scaling or normalization, simplifying preprocessing.
- Works well with both numerical and categorical input features.
- Provides transparent decision paths that help in explaining predictions.

4. Results description

Fig. 4 shows the dataset uploading interface of the system, where the selected Ayurvedic Prakriti dataset is successfully loaded and displayed within the application workspace. This step confirms proper data input before preprocessing and model training.

Fig 4: GUI for instagram engagement rate system

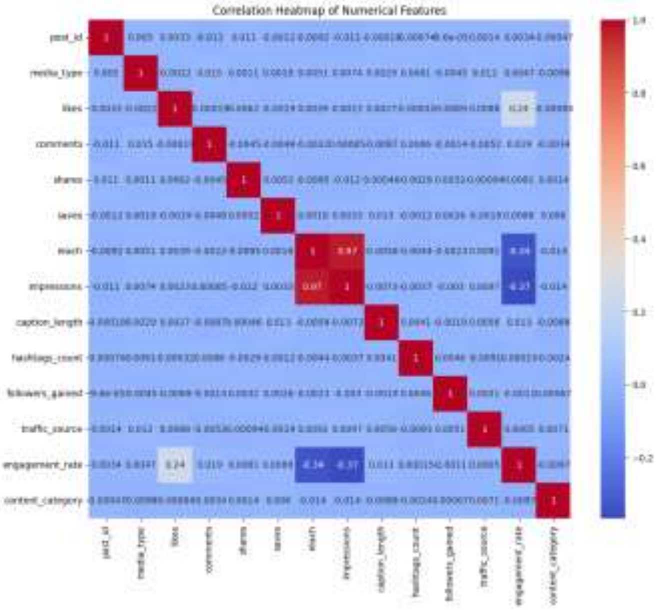
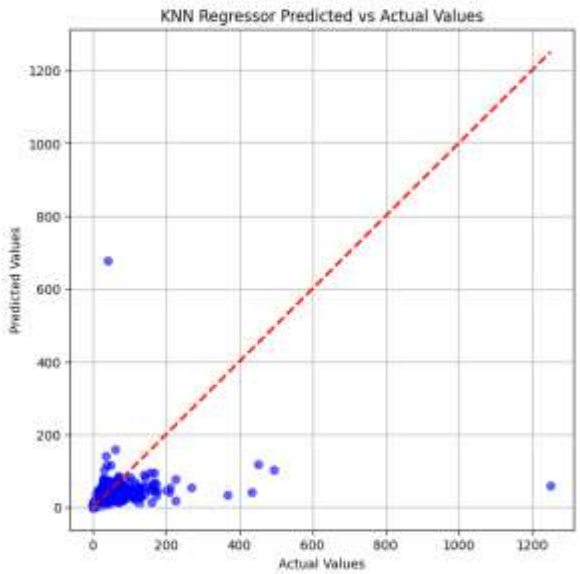
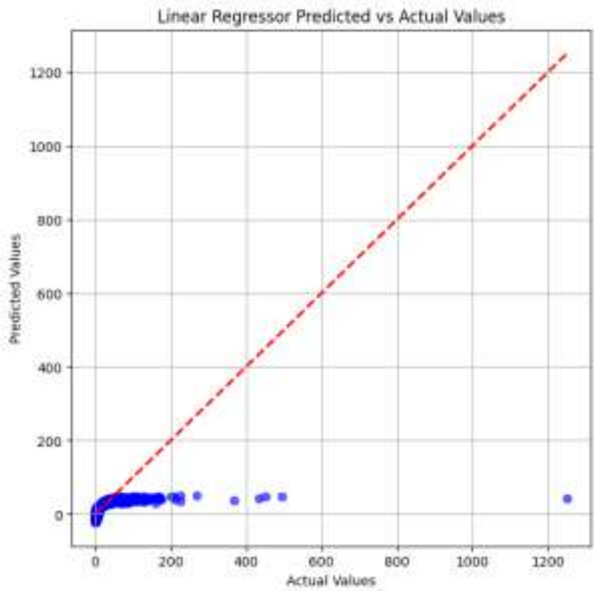


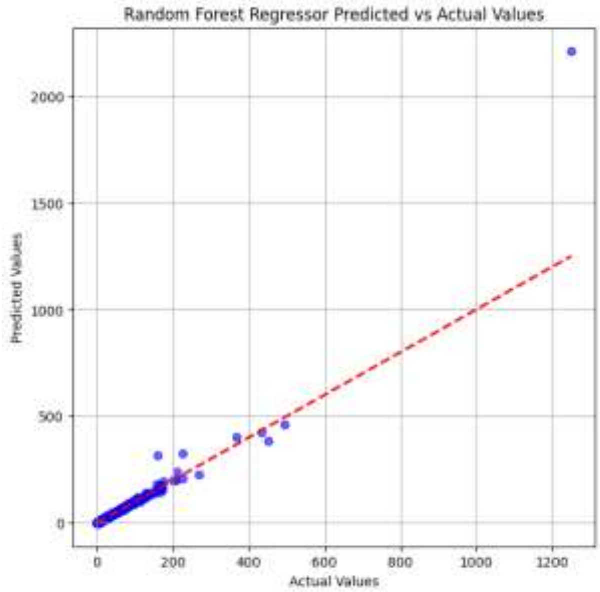
Fig 5: Correlation heatmap.

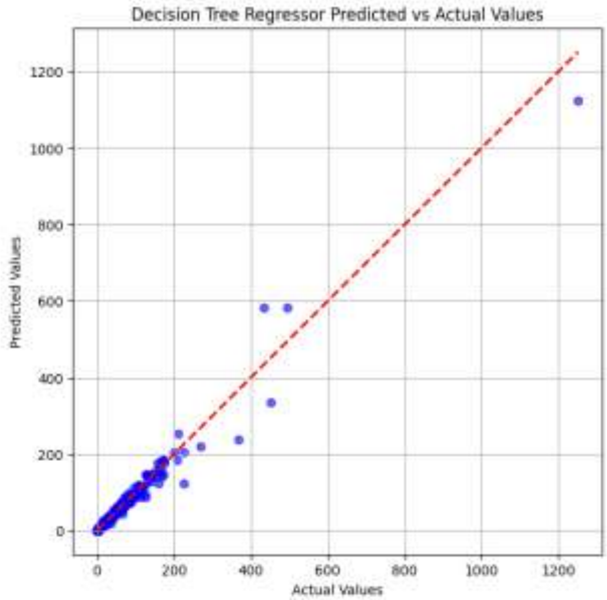
Fig 5 shows that reach and impressions have a strong positive relationship, indicating that posts viewed by more users tend to have higher display frequency. The engagement rate has a moderate positive correlation with likes, showing that likes contribute significantly to engagement. A negative correlation is observed between engagement rate, and both reach and impressions, suggesting that higher visibility does not always result in stronger interaction. Other features such as comments, shares, saves, caption length, and hashtag count exhibit very weak correlations, indicating limited direct linear influence on engagement.





(a)





(b)

(c) (d)

Fig. 6: Scatter Plot of Regression Models. (a). KNNR. (b) LR. (c). RFR. (d) DTR

Fig. 6 presents the scatter plot visualization of different regression models, including KNNR, Linear Regression (LR), Random Forest Regression (RFR), and Decision Tree Regression (DTR). The plots illustrate the relationship between actual and predicted values, highlighting variations in prediction accuracy across models. Among them, ensemble-based regression demonstrates better alignment with actual data, indicating improved predictive performance.

Table 1 presents a comparative evaluation of regression models for Instagram engagement rate prediction. Traditional models such as LR and KNNR show limited performance, with high error values and low R^2 scores, indicating poor generalization and inability to capture complex engagement patterns. The RFR demonstrates improved performance with a lower MAE and a stronger R^2 value, reflecting better handling of non-linear feature interactions. In contrast, the DTR significantly outperforms all other models, achieving the lowest error values and a very high R^2 score of 0.97. This result highlights the effectiveness of decision-based learning in modelling engagement behaviour.

Table 1: Comparative performance of regression models for engagement rate prediction

Algorithm Name	**MAE**	**MSE**	**RMSE**	**R^2**
DTR	0.95	18.97	4.35	0.97
RFR	0.59	153.64	12.39	0.77
LR	7.78	510.72	22.59	0.24
KNNR	4.25	465.34	21.57	0.31

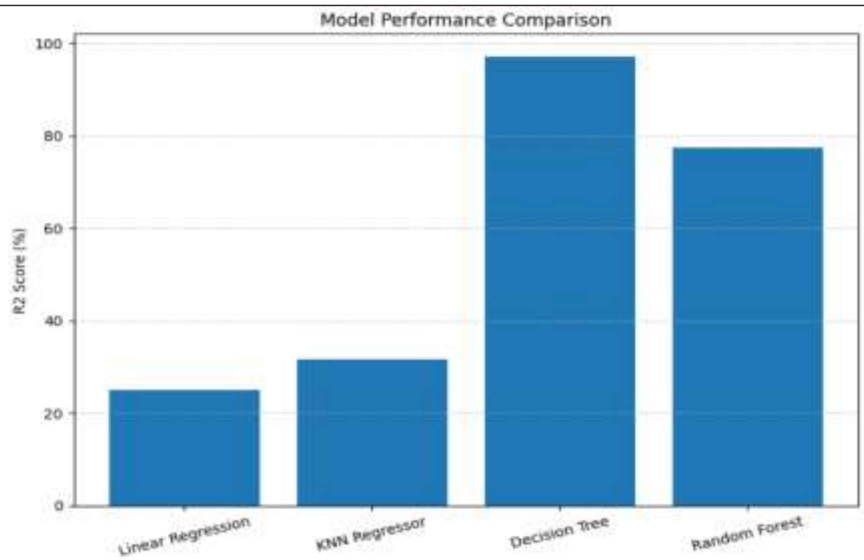


Fig. 7: Model Comparison

Fig. 7 presents the comparative performance analysis of different regression models based on the R^2 score metric. The bar chart illustrates that Linear Regression and KNN Regressor achieve relatively lower prediction accuracy, indicating limited capability in capturing complex data patterns. The Decision Tree model shows the highest R^2 score, demonstrating strong learning ability and effective modeling of nonlinear relationships. Random Forest also achieves high performance, though slightly lower than Decision Tree, highlighting the effectiveness of ensemble learning. Overall, the figure emphasizes that tree-based models provide superior predictive performance compared to traditional regression approaches.

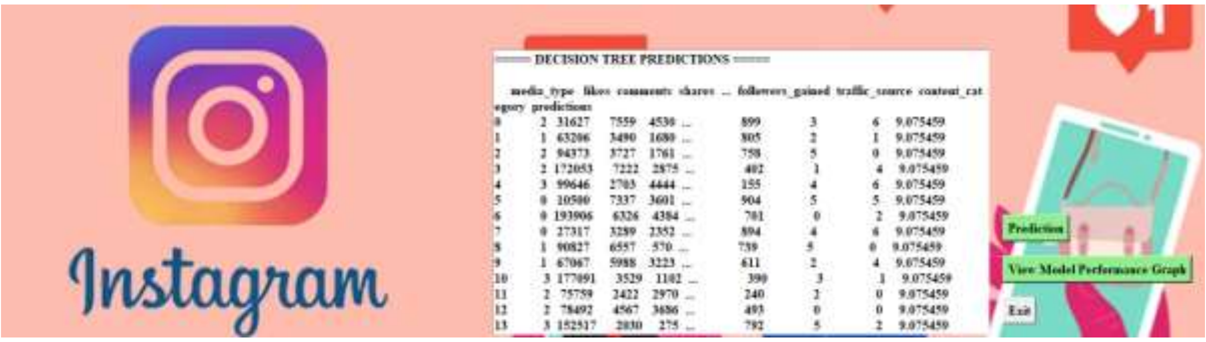


Fig. 8: Prediction results from test data.

Fig. 8 presents the prediction results generated from the test dataset after applying the trained DTR. Each row represents a processed Instagram post with standardized feature values such as likes, comments, shares, reach, and impressions. The pred_scaled column indicates the predicted engagement rate for each test instance. Variations in predicted values reflect how different combinations of user interaction and content features influence engagement. These results demonstrate the model's ability to generalize learned patterns to unseen data.

5. CONCLUSION

This research successfully developed and evaluated a machine learning-based framework for predicting Instagram engagement rates using user and content-related features. The study incorporated a comprehensive workflow consisting of data collection, preprocessing, exploratory data analysis, feature scaling, model training, performance evaluation, and engagement rate prediction. By leveraging historical Instagram analytics data, the proposed system effectively identified patterns and relationships that influence user engagement. Multiple regression algorithms, including Linear Regression (LR), K-Nearest Neighbors Regressor (KNNR), Random Forest Regressor (RFR), and Decision Tree Regressor (DTR), were implemented and compared using standard evaluation metrics such as MAE, MSE, RMSE, and R^2 Score. Experimental results demonstrated that tree-based models outperformed traditional regression approaches in capturing the complex and non-linear nature of Instagram engagement behavior. Among all the models, the Decision Tree Regressor achieved the highest predictive performance with an R^2 score of approximately 0.97, indicating excellent accuracy and reliability. The findings confirm that machine learning techniques can provide valuable insights into social media engagement dynamics and significantly enhance prediction capabilities compared to conventional analytical methods. The proposed system offers a scalable, interpretable, and data-driven solution that can assist businesses, influencers, marketers, and content creators in optimizing content strategies and improving audience engagement. Overall, this research highlights the potential of machine learning in social media analytics and establishes a strong foundation for future advancements in engagement prediction and digital marketing intelligence.

REFERENCE

[1] Eckstein, M. (2024, January 29). Social Media Engagement: Why It's Important and How to Do It Well - the Buffer Blog. Buffer Library. https://buffer.com/library/social-media-engagement

[2] Emmert-Streib, F., & Dehmer, M. (2019). High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection. Machine Learning and Knowledge Extraction, 1(1), 359–383. https://doi.org/10.3390/make1010021

[3] Yiu, T. (2019, June 12). Understanding Random Forest. Medium; Towards Data Science. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[4] Zote, J. (2024, September 26). Instagram engagement rate: How to calculate yours in 2024. Sprout Social. https://sproutsocial.com/insights/instagram-engagement-rate

[5] Recupero, D.; Nuzzolese, A.; Consoli, S.; Presutti, V.; Peroni, S.; Mongiovi, M. Extracting knowledge from text using SHELDON, a semantic holistic Framework for LinkeD ONtology data. In Proceedings of the 24th International Conference on World Wide Web, Florence, Italy, 18–22 May 2015; pp. 235–238.

[6] Consoli, S.; Recupero, D. Using FRED for named entity resolution, linking and typing for knowledge base population. Commun. Comput. Inf. Sci. 2015, 548, 40–50.

[7] Dridi, A.; Reforgiato Recupero, D. Leveraging semantics for sentiment polarity detection in social media. Int. J. Mach. Learn. Cybern. 2019, 10, 2045–2055.

[8] Carta, S.; Corriga, A.; Ferreira, A.; Podda, A.S.; Recupero, D.R. A multi-layer and multi-ensemble stock trader using deep learning and deep reinforcement learning. Appl. Intell. 2020, 1–17.

[9] Barra, S.; Carta, S.M.; Corriga, A.; Podda, A.S.; Recupero, D.R. Deep learning and time series-to-image encoding for financial forecasting. IEEE/CAA J. Autom. Sin. 2020, 7, 683–692.

[10] Padiya, V., Shah, D., Dave, T., Karani, R. (2025). Classification of Instagram Users and Prediction of Engagement Rates Using Machine Learning. In: Joshi, A., Ragel, R., Mahmud, M., Kartik, S. (eds) ICT: Applications and Social Interfaces. ICTCS 2024. Lecture Notes in Networks and Systems, vol 1323. Springer, Singapore. https://doi.org/10.1007/978-981-96-4139-0_30

[11] Son, H., & Park, Y. E. (2024). Predicting user engagement with textual, visual, and social media features for online travel agencies' Instagram post: evidence from machine learning. Current Issues in Tourism, 27(22), 3608–3622. https://doi.org/10.1080/13683500.2023.2278087

[12] R. L. H. Yew, S. B. Suhaidi, P. Seewoochurn and V. K. Sevamalai, "Social Network Influencers' Engagement Rate Algorithm Using Instagram Data," 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), Subang Jaya, Malaysia, 2018, pp. 1-8, doi: https://doi.org/10.1109/ICACCAF.2018.877675

[13] AlAnezi, M., Almutairy, M. (2021). Influencer Engagement Rate Under Scalable Machine Learning Approaches. In: Meiselwitz, G. (eds) Social Computing and social media: Applications in Marketing, Learning, and Health. HCII 2021. Lecture Notes in Computer Science (), vol 12775. Springer, Cham. https://doi.org/10.1007/978-3-030-77685-5_1

[14] P. Marimuthu, A. Garg, A. Rao and A. Thakur, "Evaluation of Instagram Performance Metrics on Technology and Innovation Hashtags," 2025 6th International Conference for Emerging Technology (INCET), BELGAUM, India, 2025, pp. 1-5, doi: https://doi.org/10.1109/INCET64471.2025.11140813.

[15] Mazloom M., Rietveld R., Rudinac S., Worring M., and Dolen W., "Multimodal Popularity Prediction of Brand-Related Social Media Posts," in Proceedings of the 24th ACM international Conference on Multimedia, Amsterdam, pp. 197- 201, 2016.