Using Machine Learning to Predict Instagram Engagement Rate

Albertus Santoso
Computer Science Department
School of Computing and Creative Arts
Bina Nusantara University,
Jakarta, Indonesia 11480
albertus.santoso@binus.ac.id

Gabriel Anderson
Computer Science Department
School of Computing and Creative Arts
Bina Nusantara University,
Jakarta, Indonesia 11480
gabriel.anderson@binus.ac.id

Rafael Anderson
Computer Science Department
School of Computing and Creative Arts
Bina Nusantara University,
Jakarta, Indonesia 11480
rafael.anderson@binus.ac.id

Nunung Nurul Qomariyah
Computer Science Department
School of Computing and Creative Arts
Bina Nusantara University
Jakarta, Indonesia 11480
nunung.qomariyah@binus.edu

Raymond Bahana
Computer Science Department
School of Computing and Creative Arts
Bina Nusantara University
Jakarta, Indonesia 11480
rbahana@binus.edu

***Abstract*—Instagram has become a tool used by marketers and influencers to engage with their audiences, with the intention of increasing popularity or promoting a product or service. Understanding trends and patterns is crucial in boosting the engagement rate received, which people may struggle with. This study aims to solve this issue by applying machine learning models like Gradient Boosting Regressor, K-Nearest Neighbors, Random Forest, XGBoost, and CatBoost. With the application of machine learning models, hidden patterns or trends that boosts the engagement rate may be discovered. From these models, XGBoost is the best performing model, giving an accuracy of (0.94), with an approximate root mean squared error of (1.26), indicating its proficiency in predicting the engagement rate values. Additionally, it was found out that the engagement rate based on the last 60 days is the most influential factor over the overall engagement rate. These results offer valuable insights for marketers or Instagram influencers trying to optimize engagement from potential audiences.**

***Index Terms*—Engagement Rate, Instagram Accounts, Machine Learning, Gradient Boost, K-Nearest Neighbors, Random Forest, XGBoost, CatBoost**

I. Introduction

In the current digital world, smartphones have become a vital tool in people's daily lives worldwide, acting as a tool for communication and as a source of entertainment [1]. As a result, social media has gained popularity, which results in the amount of social media accounts across various platforms to increase rapidly [2]. In 2020, there is an estimate of 3.6 billion active users on all social media platforms, and this number is expected to continue to grow in the future [3]. Due to the increasing popularity of social networks, it has led to an increase in the number of influencers. These influencers gain a significant amount of audience, boosting the engagement rate received by their accounts on the platform. Due to the significant count of users engaging with their content, influencers have become a powerful tool for marketers and brands to effectively reach their target audience and promote their products or services [4]. In fact, approximately 75 percent of marketers have the intention of being involved in influencer marketing, where spending related to influencer marketing reached a total of 16.4 billion dollars by the end of 2022 [5]. These spending have increased significantly to 35 billion dollars in 2024 and the value is predicted to continue increasing by 10 percent annually until 2029 [6].

Instagram is one of the most popular social media platforms currently, with 500 million users daily and over 1 billion users in total. The frequency of media sharing such as photos and videos in Instagram is higher compared to other social media platforms, averaging 95 million photos and videos shared daily [7]. Furthermore, influencers are able to engage with their audience by liking and replying the comments received, or messaging their viewers directly, enabling them to engage with their audience effectively [8]. The frequent interactions between users is also one of the factors why it is one of the most common platforms used by marketers or influencers to promote products and services, as it is highly effective in increasing brand awareness and exposure.

However, finding which Instagram accounts receive the highest overall rate of engagement from users is a challenge faced by most brands trying to find a suitable promoter for their brand. Based on a report, it found that marketers investing in influencer marketing receive disappointing results, having little to no benefit received from the campaign [5]. It is crucial to find the number of active users proportional to the number

979-8-3315-2768-6/24/$31.00 © 2024 IEEE

of followers the account has. As brands would prefer ”quality” users rather than ”quantity” users. For example, brands would rather find Instagram accounts a higher proportion of users that are active and genuinely interested in the influencer’s content rather than an account with more likes but a lower engagement rate indicating inactive or disinterested users. Interested users will talk and share the product, providing additional exposure to the promoted product or service. This is a crucial factor that must be considered when choosing an influencer as a sponsor.

Due to the strong correlation between the size of an influencer with the cost of sponsor, choosing the right influencer may result in better engagement rate, while paying lower costs. For example, sponsoring a smaller influencer will likely cost less than a larger influencer. By finding one that has decent engagement rate, the product promotion may have a more significant impact on its audience, while costing less for the sponsor.

II. Related Work

Based on a study conducted by Trunfio and Rossi [9], social media engagement has a multidimensional and polysemic nature, requiring complex models to analyze it. The study suggests the usage of the COBRA model (Consumer Online Brand Related Activities) as a tool to analyze the engagement on social media. The COBRA model analyzes all 3 dimensions of actions from the behavioral perspective, which are consumptions, contributions, and creation. Basically, it applies different weights for different social media platforms, as some social media platforms can remove or hide the exact number of social media metrics, including likes, comments, and subscribers, which can result in inaccuracies [10].

A research by Shahzad et al. [11] explored the aspects of influencers, and how those aspects affect their engagement rate. The engagement of the follower is determined by three critical factors, including content, source, and psychological characteristics, where source and psychological characteristics receive limited attention during engagement analysis, compared to content characteristics. To resolve this issue, the research proposes a model that is supported by both Elaboration Likelihood Model (ELM) and Dual Process Theory. The ELM is able to highlight the content’s quality, and the attractiveness of the source. On the other hand, Dual Process Theory focuses more on psychological aspects. The research explains how consumers switch between automatic and reflective processing when encountering influencer content.

Another study by Tan and Lim [12] uses past studies, which circulate about engagement rates on social media, finding patterns from these studies. One of the findings states that brand awareness has a significant impact on the impression rate and reach. The study suggests understanding of social media metrics and analytics based on one’s business. Additionally, providing consumers with content they want is crucial, and less of the brand’s promotion. Lastly, it states that analyzing the optimal way of calculating social media metrics to measure a business’ brand awareness, and observing how social media is adopted by different industries are important, and must be considered. Lastly, a study by Putranto et al. [13] revolves around the studying of accounts of MSME actors, where it was found that most MSME weren’t able to optimize their accounts in attracting engagement. It stated that the engagement level is crucial, as it tells how much influence and impact an account has on its followers.

Our study aims to approach these issues by implementing a unique approach, machine learning. Unlike some of the studies, this study solely focuses on the engagement rate on the Instagram platform. By performing analysis on the quantitative aspects, new patterns can be discovered. Furthermore, implementing various machine learning models with different levels of complexities can result in a better engagement rate analysis. Hopefully, the insights discovered from this study can be helpful, providing marketers with a reliable method in increasing the engagement rate of their Instagram account. Machine learning models like Gradient Boosting Regressor, K-Nearest Neighbors, Random Forest, XGBoost, and CatBoost, are utilized.

III. Methodology

A. Dataset Overview and Description

This study’s dataset is mainly sourced from scraping, which is done by utilizing a python package called Instaloader. Instaloader is able to access the information of an Instagram account, post details, etc. A total of 743 rows are obtained from scraping for our dataset. Aside from data scraping, some parts of the data are obtained from Kaggle[1], which contains the data regarding the top 200 most followed Instagram accounts worldwide. The dataset from Kaggle were manually scraped by the author from the Socialbook website [14]. The size of the data after cleaning is (922, 9), with 922 records and 9 attributes.

The accounts from the dataset can be separated into several categories. This includes the origin of the account which is either Global or Indonesia, followed by their account size based on the followers, including mega, macro, micro, and nano. These attributes and their descriptions are shown in the following Table I.

The attributes of the dataset, along with their description and sample values are summarized in Table II.

Several Python libraries are used to collect, clean, and analyze the dataset. It is also used for data visualization. The details of the libraries used are shown in Table III.

B. Dataset Preprocessing

To process our data to be readable and usable, we first need to format the csv files. Since the accounts scraped and the data

[1] See more at: https://www.kaggle.com/datasets/syedjaferk/top-200-instagrammers-data-cleaned

TABLE I
ACCOUNT ORIGIN

Attribute	Description
Global	Accounts across the world
Indonesia	Accounts from Indonesia
Mega	More than 1M Followers
Macro	100k - 1M Followers
Micro	10k - 100k Followers
Nano	1k - 10k Followers

TABLE II
MERGED INSTAGRAM ACCOUNTS DATASET

Attribute	Description	Sample Data
Username	Username of the account	cristiano
Followers	Follower count of the account	47,352,198
Average Likes	The average likes of the account	458,723.76
Average Comments	The average comments of the account	4,422.62
Average Engagement Rate	Value calculated from the average likes, comments, and followers	0.9781
Engagement Rate (60 Days)	Same as the average engagement rate, but only containing data from posts posted the last 60 days	0.4896
Posting Frequency (60 Days)	Number of posts posted (60 Days)	132
Posts (Image)	Number of posts of the image type (60 Days)	9
Posts (Video)	Number of posts of the video type (60 Days)	34
Posts (Carousel)	Number of posts of the carousel type (60 Days)	89
Average Hashtags / Post (60 Days)	Average number of hashtags used for posts posted last 60 days	1.0

we got from Kaggle have their respective csv files, the datasets need to be formatted individually. The data from Kaggle are cleaned by removing unnecessary columns that are not used in the analysis. It is also added with additional attributes to match the structure of the scraped data. As an example, data from the last 60 days are added to the Kaggle data. Then, both csv files are merged together, and the outliers are removed from the data by using the Z-score method. The detailed process can be seen in Fig. 1.

TABLE III
LIBRARIES USED IN PYTHON

Library	Description
instaloader	Tool to scrape and download instagram data
pandas	Library to clean, transform, and analyze datasets
numpy	Numerical computing library to fill missing values
matplotlib	Library used for data visualization
seaborn	Library used for data visualization

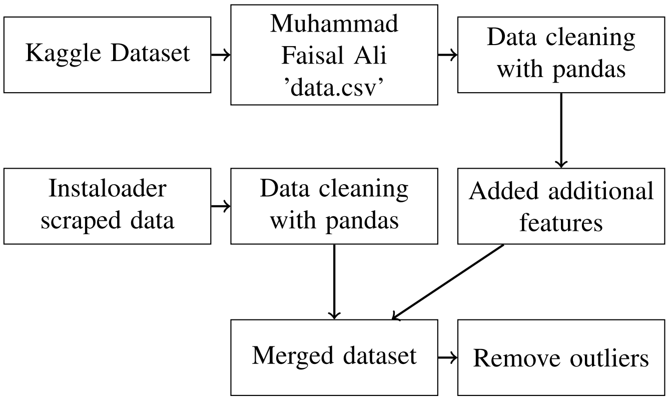
Kaggle Dataset
Muhammad Faisal Ali 'data.csv'
Data cleaning with pandas
Instaloader scraped data
Data cleaning with pandas
Added additional features
Merged dataset
Remove outliers

Fig. 1. Flowchart illustrating the process of data preprocessing.

C. Model and Techniques

We used five supervised machine learning models to analyze the dataset:

- **Gradient Boosting Regressor** has the ability to create a connection with a statistical framework. It is able to provide any necessary justifications towards the hyperparameters used, for improvement in model performance. [15]. It is used for its high efficiency and ability to handle complex relationships.
- **XGBoost** is an improved version of the gradient boosting model, with it being scalable. It is suitable for use when you want to increase the accuracy of a boosting model in a short period of time [16]. Similar to the Gradient Boosting Regressor, it was chosen for its high efficiency in dealing with complex relationships and high accuracy.
- **Random Forest** contains decision trees, where each tree trains on a random part of the data, and these trees are distributed throughout the forest [17]. It was chosen due to its capabilities in analyzing the data effectively regardless of the data being linear or not.
- **CatBoost** is also stated to be the refined version of gradient boosting, where it is able to deal with high cardinality categorial variables. Furthermore, it also has the ability to greedily choose feature combinations that may improve the performance of the model [18]. It was chosen for its high efficiency and ease of use.
- **K-Nearest Neighbor (KNN)** finds the distance between the query and current examples in the data, selecting the k-nearest neighbors with the smallest distance and averages their labels [19]. It was chosen for its effectiveness in capturing local patterns in the data.

The label used for this dataset is 'Average Engagement Rate', which translates to the average engagement rate for a post. All the features used to predict the label are continuous variables, we used 'Followers', 'Average Likes', 'Average Comments', 'Engagement Rate (60 Days)', 'Posting Frequency (60 Days)', 'Posts (Image)', 'Posts (Video)', 'Posts (Carousel)'. The description of features used in the dataset is shown in Table IV.

TABLE IV
FEATURES USED IN THE MACHINE LEARNING MODELS

Features	Description
Followers	Total Followers count of accounts
Average Likes	Average likes of accounts
Average Comments	Average commments of accounts
Engagement Rate(60days)	Average engagement of post in the last 60 days
Posting Frequency(60days)	Number of post in the last 60 days
Post(image)	Image type post in the last 60 days
Post(Video)	Video type post in the last 60 days
Post(Carousel)	Carousel type post in the last 60 days

D. Evaluation Techniques

To evaluate which machine learning model produces the best results, several evaluation techniques are used. The evaluation technique includes Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R2 score.

1) Mean Squared Error (MSE): Mean squared error is used as a representation of how well a model reproduces reality, which is obtained from compressed data used in training and resulted in prediction. However, by compressing the data into one value, it provides few insights on which data are helpful and which are useless [20].

2) Root Mean Squared Error (RMSE): Root mean squared error (RMSE) is a standard statistical technique used in model evaluation. It quantifies the average magnitude of error between the predicted and actual values, with the error being averaged and square-rooted [21].

3) R2 score: The R2 score is used to represent the percentage of the variation explained compared to the total variation. It helps to measure the spread of data points around the regression line [22].

These evaluation techniques provide a balanced evaluation and insight of model performance, capturing the error values and the proportion of the data explained.

IV. RESULTS AND DISCUSSION

As shown in the following Fig. 2, the XGBoost model performs exceptionally well in predicting engagement rates. Most of the predicted data points were closely aligned with the red line representing the perfect prediction line, signifying a high accuracy and minimal error in the model's prediction.

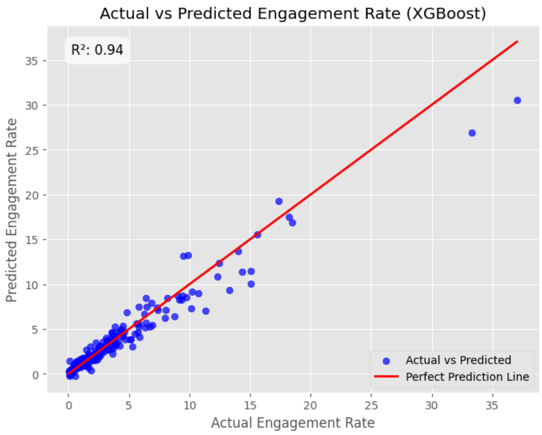

Fig. 2. XGBoost Graph

TABLE V
MODELS PERFORMANCE

Models	MSE	RMSE	R2
XGBoost	1.60	1.26	0.94
Gradient Boosting Regressor	1.93	1.39	0.92
CatBoost	2.38	1.54	0.91
KNN	4.52	2.13	0.82
Random Forest	6.83	2.61	0.73

From the results shown in TABLE V, we can order the performance of the models from best to worst based on their evaluation metrics, starting from XGBoost, Gradient Boosting Regressor, CatBoost, KNN, Random Forest. The XGBoost model performs the best and most accurate in all evaluation aspects compared to other models with an MSE of 1.60, RMSE of 1.26, and R2 score of 0.94. On the other hand, the Random Forest model performed poorly compared to the other models with an MSE of 6.83, an RMSE of 2.61, and an R2 of 0.73.

The results show that the gradient boosting models (Gradient Boosting Regressor, XGBoost, and CatBoost) performed significantly better compared to the other models, indicating that the gradient boosting models are suitable for this task. From this insight, we can conclude that gradient boosting models performed superior mainly due to their ability to capture complex, non-linear relationships, and feature interactions. Other models such as KNN may struggle with noisy data and Random Forest failing to capture subtle patterns which caused them to be outperformed by the gradient boosting models.

From our analysis after cleaning the dataset as shown in Fig. 3, there is a strong positive correlation between average likes and average comments, indicating that posts with higher likes tend to have more comments. Furthermore, there is a moderate positive correlation between the engagement rate in the last 60 days and average engagement rate, indicating consistency between the short-term and long-term trends.

One of the surprising findings was that on median, there was a large difference in engagement between the video and carousel post-type rate with the image post-type, where it seemed that the image post-type struggled greatly to receive engagement, which can be seen in Fig. 4. Another interesting

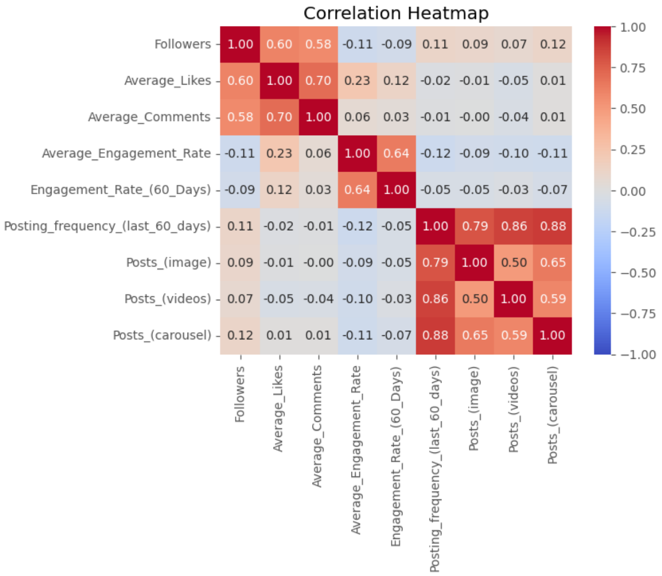

Fig. 3. Correlation Matrix

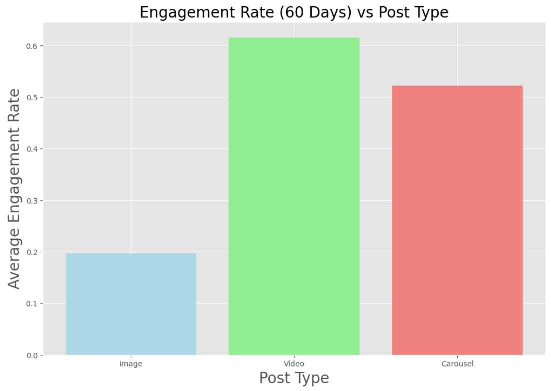

Fig. 4. Engagement Rate (60 Days) vs Post Type

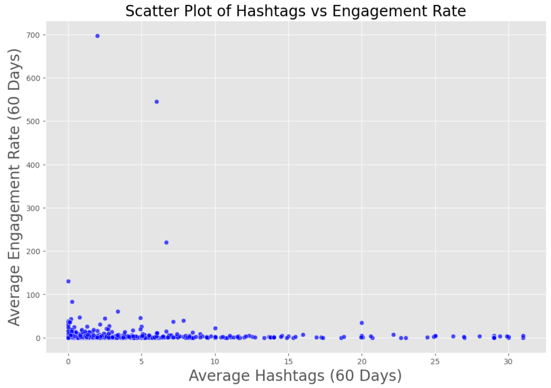

Fig. 5. Engagement Rate (60 Days) vs Average Hashtags Used

insight is that the use of hashtags has little to no impact towards the engagement rate. Most accounts have a relatively consistent engagement rate regardless of the number of hashtags used, with a few accounts obtaining a higher engagement rate than those with lots of hashtags used, which can be seen on Fig. 5. Another surprising finding was that the nano-type accounts received the highest engagement rate on the median, followed by macro- and micro-type accounts, whereas the mega-type accounts received the lowest engagement rate. This indicates that accounts with a smaller following are more likely to attract a higher rate of new users, which may be beneficial for brands that are looking for strategic methods to attract new consumers and be cost-efficient.

V. Conclusion and Future Work

To conclude, this study aims to predict the engagement rate of posts posted by an Instagram account. Engagement rate is a key metric used in evaluating the effectiveness of influencer marketing strategies. Five machine learning models were used in this study including XGBoost, CatBoost, Gradient Boosting Regressor, K-Nearest Neighbor (KNN), and Random Forest. The performance of these models are then compared using evaluation techniques such as the R2 score, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

The XGBoost model achieved the best results with a root mean squared error of 1.26, indicating that the predicted value is close to the actual value. On the other hand, the Random Forest model performed relatively poor compared to other models with a root mean squared error of 2.61, showing a significant gap compared to the XGBoost model. This might be caused by the model failing to identify subtle patterns. The gradient boosting models (XGBoost, CatBoost, and Gradient Boosting Regressor) performed better compared to other models, making them the most suitable type for this research. This is likely due to their ability to capture complex, non-linear relationships and feature interactions. Gradient Boosting models iteratively correct errors through boosting, offering higher accuracy and better generalization.

These findings are particularly relevant for brands and influencers who want to optimize their marketing strategies on Instagram. Smaller influencers with a high engagement rate can be cost effective while having a good engagement rate, making them a choice for brands with limited budgets.

For our future work, we would like to first add more data, as for this research we have only used the data from one social media platform which is Instagram. We would like to add other social media platforms such as Tiktok or X. Next, we would like to add the data volume as our data after cleaning is 922 accounts and as we all know the more the data the better the outcome. We would also like to find a better and more reliable scraper as Instaloader is unstable and costs us time that we could have used to add more data. Predicting the engagement rate in real time is needed for our research to have a better accuracy as followers, likes, and comments change from time to time. We would implement the hashtag dictionary for machine learning as for the research as we did not implement it.

Supplementary Codes

All the codes used in this paper can be accessed through the following link: https://github.com/gamakagami/FoDS-FinalProject

References

[1] A. K. Nuhel, "Evolution of smartphone," October 2021.

[2] A. A. Arman and A. P. Sidik, "Measurement of engagement rate in instagram (case study: Instagram indonesian government ministry and institutions)," in *2019 International Conference on ICT for Smart Society (ICISS)*, vol. 7, pp. 1–6, IEEE, 2019.

[3] C. Zachlod, O. Samuel, A. Ochsner, and S. Werthmüller, "Analytics of social media data – state of characteristics and application," *Journal of Business Research*, vol. 144, pp. 1064–1076, 2022.

[4] M. Pretel-Jiménez, J.-L. del Olmo, and C. Ruíz-Viñals, "The engagement of literary influencers with their followers on instagram: Bookstagrammers' content and strategy," *Revista Mediterránea de Comunicación/Mediterranean Journal of Communication*, vol. 15, no. 1, pp. 305–321, 2024.

[5] F. F. Leung, F. F. Gu, Y. Li, J. Z. Zhang, and R. W. Palmatier, "Influencer marketing effectiveness," *Journal of Marketing*, vol. 86, 2022.

[6] K. Spörl-Wang, F. Krause, and S. Henkel, "Predictors of social media influencer marketing effectiveness: A comprehensive literature review and meta-analysis," *Journal of Business Research*, vol. 186, p. 114991, 2025.

[7] P. Bellavista, L. Foschini, and N. Ghiselli, "Analysis of growth strategies in social media: The instagram use case," in *2019 IEEE 24th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, University of Bologna, 2019.

[8] M. U. Chaudhary, "Impact of instagram as a tool of social media marketing," *Media and Communication Review*, vol. 1, no. 1, pp. 17–29, 2021.

[9] M. Trunfio and S. Rossi, "Conceptualising and measuring social media engagement: A systematic literature review," *Italian Journal of Marketing*, vol. 2021, pp. 267–292, 2021.

[10] B. Schivinski, G. Christodoulides, and D. Dabrowski, "Measuring consumers' engagement with brand-related social-media content: Development and validation of a scale that identifies levels of social-media engagement with brands," *Journal of Advertising Research*, vol. 56, no. 1, pp. 1–18, 2016.

[11] A. Shahzad, H. Rashid, A. Nadeem, M. Bilal, and W. Ahmad, "Social media influencer marketing: Exploring the dynamics of follower engagement," *Journal of Policy Research*, no. 4, pp. 1–8, 2023.

[12] W. B. Tan and T. Lim, "A critical review on engagement rate and pattern on social media sites," in *Proceedings of the International Conference on Digital Transformation and Applications (ICDXA 2020)*, TARUMT, TARUC, January 2020.

[13] H. A. Putranto, D. P. S. Setyohadi, T. Rizaldi, E. S. J. Atmadji, H. Y. Riskiawan, and I. H. Nuryanto, "Measurement of engagement rate on instagram for business marketing (case study msme of dowry in jember)," 2022. Available: https://www.researchgate.net/publication/366171806_Measurement_of_Engagement_Rate_on_Instagram_for_Business_Marketing_Case_Study_MSME_of_Dowry_in_Jember.

[14] Socialbook, "Top 200 instagrammers," n.d. Available: https://socialbook.io/instagram-channel-rank/top-200-instagrammers.

[15] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neurorobotics*, vol. 7, December 2013.

[16] S. Malik, R. Harode, and A. Singh, "Xgboost: A deep dive into boosting (introduction documentation)," February 2020.

[17] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," in *Ensemble Machine Learning* (C. Zhang and Y. Ma, eds.), pp. 157–175, New York, NY: Springer, 2012.

[18] J. T. Hancock and T. M. Khoshgoftaar, "Catboost for big data: an interdisciplinary review," *Journal of Big Data*, vol. 7, no. 1, p. 94, 2020.

[19] O. Harrison, "Machine learning basics with the k-nearest neighbors algorithm," 2019. Available: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761.

[20] T. O. Hodson, T. M. Over, and S. S. Foks, "Mean squared error, deconstructed," *Journal of Advances in Modeling Earth Systems*, vol. 13, p. e2021MS002681, 2021. First published: 23 November 2021.

[21] T. Hodson, "Root-mean-square error (rmse) or mean absolute error (mae): when to use them or not," *Geoscientific Model Development*, vol. 15, no. 14, pp. 5481–5487, 2022.

[22] D. Figueiredo, J. Silva, and E. C. R. Júnior, "R2 all about?," *Leviathan*, vol. 3, pp. 60–68, Nov 2011.