Buletin Ilmiah Informatika Teknologi
Vol 4, No 1, September, Hal. 25 - 32
ISSN 2962-0945 (media online)
DOI 10.58369/biit.v2i3.167
https://ejurnal.amikstiekomsu.ac.id/index.php/BIIT

Uncovering Smartphone Brand Strategies through Specification-Based Clustering and Classification

Abdul Karim[1*], Andi Ernawati[2]

[1]Faculty of Science and Technology, Information Technology Study Programme, Universitas Labuhanbatu, Indonesia,
[2]Master of Information Technology, Universitas Pembangunan Panca Budi Medan, Indonesia
Email:[1]abdkarim6@gmail.com, [2]aernawati296@gmail.com
Corresponding email: Abdkarim6@gmail.com

Abstract - − In an increasingly saturated smartphone market, brand differentiation through technical specifications has become a core strategy for attracting diverse consumer segments. This study proposes a machine learning approach to uncover underlying brand strategies by leveraging smartphone specifications and market pricing across multiple regions. We utilize unsupervised clustering algorithms (K-Means, DBSCAN) to segment devices based on technical features, followed by supervised classification models (Random Forest, XGBoost) to identify and interpret brand-driven design strategies. The dataset comprises smartphones released in 2024–2025, including attributes such as RAM, camera specifications, processor type, battery capacity, and launch prices in Pakistan, India, China, USA, and Dubai. Our findings reveal distinct clusters that align with different pricing tiers and show clear brand positioning patterns. Feature importance analysis using SHAP highlights battery capacity, screen size, and processor type as critical variables influencing brand classification. This study provides valuable insights for both manufacturers and consumers in understanding competitive product strategies within the global smartphone market.

Kata Kunci: Smartphone clustering; Machine learning; Brand strategy; Specification-based analysis; Classification; SHAP.

1. INTRODUCTION

The global smartphone market is currently growing rapidly with increasingly fierce competition every year. Manufacturers are racing to deliver innovations and increasingly diverse product variants to reach various consumer segments.[1]. In such conditions, the ability of brands to differentiate themselves from competitors through technical specification strategies becomes one of the keys to winning the market [1]

Specifications such as battery capacity, camera, processor, and screen size are no longer just technical features, but also part of the brand positioning strategy in the minds of consumers [2]. For example, brands such as Apple consistently maintain a premium approach with a focus on user experience and hardware efficiency, while brands such as Xiaomi tend to target value-for-money with high specifications at mid-range prices.

Although such brand strategies have been widely observed by industry players, systematic data-driven research exploring the relationship between technical specifications and brand strategies remains limited. Most previous studies have only focused on price predictions or smartphone model classification.[3], [4], [5], without considering how the combination of specifications reflects the brand's overall strategy..

This study aims to fill this gap by adopting a machine learning-based approach. By using clustering algorithms such as K-Means [6], [7], [8], [9], [10] dan DBSCAN[11], [12], [13], This study categorizes smartphones based on similarities in technical specifications, regardless of brand name. Furthermore, classification algorithms such as Random Forest are used. [14], [15], [16] and XGBoost[17], [18] to understand how certain brands dominate or spread within specific clusters, thereby enabling conclusions to be drawn about the strategies they use.

In addition, interpretability approaches such as SHAP are used to identify the features that contribute most to differentiating strategies between brands.[19], [20], [21].[22], [23] This approach is expected to provide new insights into the smartphone competitive landscape from a specification and data perspective, and help manufacturers develop more targeted product strategies.

Thus, this study not only offers theoretical contributions in the fields of data science and technology marketing, but also strong practical relevance for the highly dynamic global smartphone industry.

2. RESEARCH METHODOLOGY

2.1 Data Set Collection

This study uses a comprehensive dataset of smartphone specifications collected from various global brands obtained from the Dataset https://www.kaggle.com/datasets/abdulmalik1518/mobiles-dataset-2025. The dataset consists of 930 rows and 19 columns, with each row representing one smartphone model released during the period 2024–2025. The data includes important attributes such as brand, screen size, RAM capacity, internal memory, camera resolution, battery capacity, operating system, price, and other technical specifications.

Buletin Ilmiah Informatika Teknologi
Vol 4, No 1, September, Hal. 25 - 32
ISSN 2962-0945 (media online)
DOI 10.58369/biit.v2i3.167
https://ejurnal.amikstiekomsu.ac.id/index.php/BIIT

Most features in the dataset are numerical, such as screen size (inches), battery capacity (mAh), and camera megapixels, while the main categorical feature is the brand, which is the target of classification. All numerical attributes have been converted to a uniform unit, and all irrelevant or redundant data has been removed during the data cleaning stage.
The data distribution was checked to ensure there was no extreme imbalance (class imbalance) in the brand labels. Missing value analysis shows that most features have good data completeness, but for certain features such as secondary camera or NFC availability, some missing values were found and then addressed with imputation techniques. Initial descriptive statistics show significant variation between brands in terms of device performance, especially in terms of price, battery capacity, and memory size.
This dataset provides a strong foundation for machine learning-based analysis, particularly in clustering for market segmentation and brand classification based on technical attributes. Next, the dataset was processed through a machine learning pipeline that included normalization, feature engineering, model training, and interpretation using the SHAP approach.

Table 1. Data Set

Company Name	Model Name	Mobile Weight	RAM	Front Camera	Back Camera	Processor	Battery Capacity	Screen Size	Launched Price (Pakistan)	Launched Price (India)	Launched Price (China)	Launched Price (USA)	Launched Price (Dubai)	Launched Year
Apple	iPhone 16 128GB	174g	6GB	12MP	48MP	A17 Bionic	3,600mAh	6.1 inches	0.354839	0.275093	0.302857	0.018208	0.231481	2024
Apple	iPhone 16 256GB	174g	6GB	12MP	48MP	A17 Bionic	3,600mAh	6.1 inches	0.371817	0.293680	0.320000	0.019472	0.250000	2024
Apple	iPhone 16 512GB	174g	6GB	12MP	48MP	A17 Bionic	3,600mAh	6.1 inches	0.388795	0.312268	0.342857	0.020737	0.268519	2024
Apple	iPhone 16 Plus 128GB	203g	6GB	12MP	48MP	A17 Bionic	4,200mAh	6.7 inches	0.397284	0.312268	0.325714	0.020737	0.268519	2024
….	….	….	….	….	….	….	….	….	….	….	….	….	….	….
Apple	iPhone 16 Plus 256GB	203g	6GB	12MP	48MP	A17 Bionic	4,200mAh	6.7 inches	0.414261	0.330855	0.342857	0.022001	0.287037	2024

2.2. Proposed Method

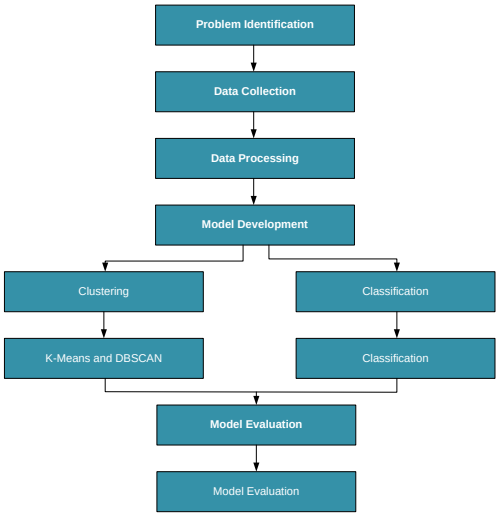
This study proposes an integrated approach to understand product specification strategies and smartphone brand classification through a combination of clustering and classification techniques based on machine learning. This approach is designed to address two main objectives: first, to identify product segmentation based on specification similarity through unsupervised learning techniques; second, to predict smartphone brands using a supervised learning model based on technical specifications.

In the clustering stage, the K-Mean algorithm is used.[24] and DBSCAN[25], [26]. The K-Means algorithm was chosen for its efficiency in clustering high-dimensional data, while DBSCAN was used as a complement to handle data with irregular distributions and possible outliers. The clustering results are expected to describe market segmentation patterns based on technical specifications such as screen size, RAM, internal storage, battery capacity, and camera resolution.
Meanwhile, for smartphone brand classification, two algorithms were applied: Random Forest [27] and XGBoost[28]. Random Forest offers model stability and resistance to overfitting [29], Meanwhile, XGBoost was chosen for its ability to provide high accuracy through efficient boosting and regularization mechanisms[30] Both models were trained using engineered and transformed numerical and categorical features.
To ensure model transparency and interpretability, this study also integrated the SHAP (SHapley Additive exPlanations) technique. SHAP provides quantitative insights into the contribution of each feature in the classification process, so it can be used to explain the influence of specific specifications on brand preferences objectively..

2.3. Research Framework

The following diagram outlines the overall research methodology for clustering using the K-Means and DBSCAN algorithms and smartphone brand classification, applying two algorithms: Random Forest and XGBoost, as well as testing the interpretability of the model using SHAP (SHapley Additive exPlanations) as follows

Buletin Ilmiah Informatika Teknologi
Vol 4, No 1, September, Hal. 25 - 32
ISSN 2962-0945 (media online)
DOI 10.58369/biit.v2i3.167
https://ejurnal.amikstiekomsu.ac.id/index.php/BIIT



This study began with the problem identification stage, which involved gaining an in-depth understanding of the product differentiation strategies of various smartphone brands through their technical specifications. Once the problem had been formulated, the data collection stage began, which involved compiling a dataset of smartphone specifications from various reliable sources. This dataset then went through a data processing stage, including data cleaning, missing value handling, normalization, and attribute transformation to make it ready for modeling.

The next stage was model development, which was divided into two main paths: clustering and classification. In the clustering path, the K-Means and DBSCAN algorithms were used to group smartphone products based on specification similarities, without considering brand labels. This aims to uncover hidden patterns or market segmentation.

Meanwhile, the classification path uses Random Forest and XGBoost algorithms to predict smartphone brands based on technical specification inputs. The selection of these two algorithms is based on their high performance in multi-class classification and their ability to handle tabular data efficiently.

After the model is developed, model performance evaluation is conducted to assess accuracy, precision, and other metrics. The final step is model interpretation using the SHAP (SHapley Additive exPlanations) approach to explain the contribution of each feature in the model's decision-making process. This approach is important to ensure that the model is not only accurate but also understandable to non-technical stakeholders..

3. RESULTS AND DISCUSSION

3.1 Research Results

From the results of data collection and exploration, the data was adjusted to be tested using clustering with the K-Means clustering and DBSCAN clustering algorithms to obtain the following patterns.

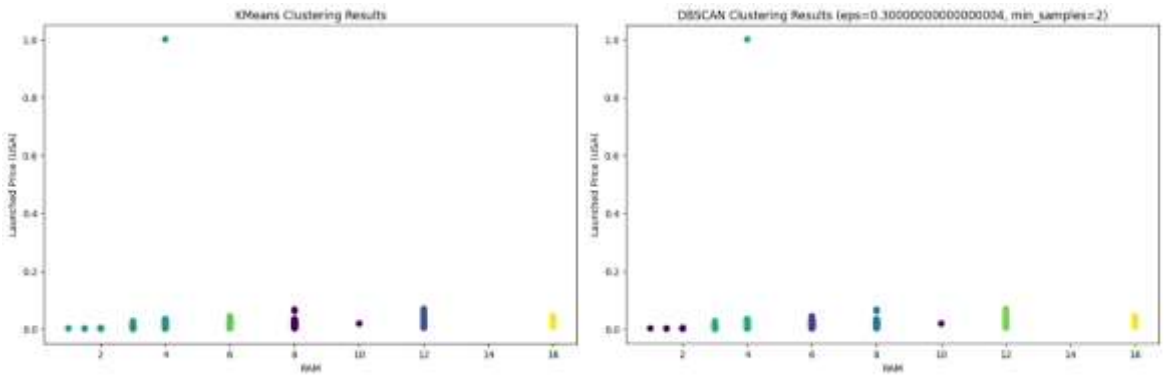


Figure 2. Clustering graph

Based on the visualization results of clustering using the KMeans and DBSCAN algorithms on RAM and launch price (USA) data, it can be seen that both methods successfully grouped the data into several clusters representing device categories based on specifications and prices. In the KMeans results, the data is divided into five main clusters that are fairly neat and separate, supported by a high Silhouette Score of 0.9069, indicating that the separation between clusters

Buletin Ilmiah Informatika Teknologi
Vol 4, No 1, September, Hal. 25 - 32
ISSN 2962-0945 (media online)
DOI 10.58369/biit.v2i3.167
https://ejurnal.amikstiekomsu.ac.id/index.php/BIIT

is quite optimal and the cohesiveness within clusters is very good. Meanwhile, the DBSCAN results with the best parameters (eps = 0.3 and min_samples = 2) yielded five main clusters and one additional cluster labeled with -1, indicating the presence of data considered outliers or noise. Interestingly, DBSCAN produced a higher Silhouette Score of 0.9272, indicating that DBSCAN is capable of identifying a more natural and flexible cluster structure in the data. The second approach shows that there is a clear relationship between RAM capacity and device launch price, with cluster patterns indicating a trend that the higher the RAM, the higher the launch price, both in the US and Indian markets..

Silhouette Score: 0.9869201826233514

cluster_label	RAM	Launched Price (USA)	Launched Price (India)
0	8.006472	0.012134	0.150380
1	12.000000	0.020836	0.266899
2	3.714286	0.012023	0.089543
3	6.000000	0.010172	0.141061
4	16.000000	0.022282	0.317904

dbscan_cluster_label	RAM	Launched Price (USA)	Launched Price (India)
-1	2.600000	0.003111	0.029551
0	6.000000	0.010172	0.141061
1	8.000000	0.012114	0.190325
2	3.811111	0.012552	0.093307
3	12.000000	0.020836	0.266899
4	16.000000	0.022282	0.317904

Best eps: 0.30000000000000004, Best min_samples: 2, Best Silhouette Score: 0.9271995372096029

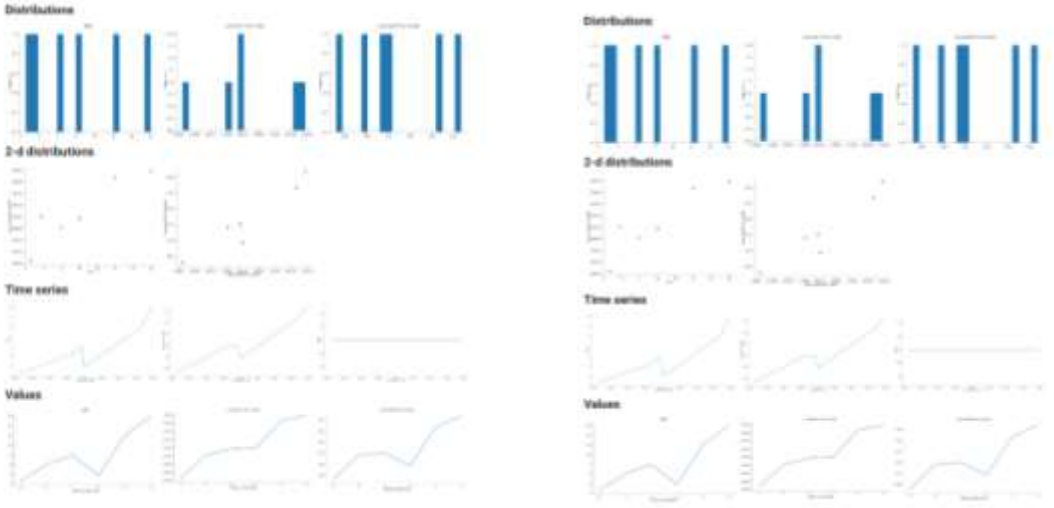


Figure 3. Graph of the method

3.2. Random Forest and XGBoost model optimization

Optimize the hyperparameters of Random Forest and XGBoost models using RandomizedSearchCV. The modeling results using the Random Forest and XGBoost algorithms show excellent performance in predicting the data used. In the Random Forest model, the best combination of hyperparameters was obtained with n_estimators of 400, min_samples_split of 2, min_samples_leaf of 2, and max_depth of 30. With this configuration, the Random Forest model was able to achieve a high accuracy score of 0.9973, reflecting the model's ability to capture complex patterns in the data. Meanwhile, the XGBoost model showed slightly better performance with an accuracy score of 0.9987, ` learning_rate `, and ` col_. The best hyperparameters for this model consist of n_estimators of 800, max_depth of 10, learning_rate of 0.1, and subsample and colsample_by_tree values of 0.8, respectively. This combination allows the XGBoost model to learn optimally from the data while maintaining a balance between bias and variance.
Overall, both models show excellent performance, but XGBoost is slightly superior to Random Forest in terms of accuracy on the data used. Therefore, XGBoost can be considered the preferred choice for prediction implementation in this scenario, especially if a more precise model is required.

3.3 Evaluation of Random Forest and XGBoost models

Evaluating the performance of the optimized Random Forest and XGBoost models using appropriate metrics, generating predictions, and visualizing the confusion matrix...

Buletin Ilmiah Informatika Teknologi
Vol 4, No 1, September, Hal. 25 - 32
ISSN 2962-0945 (media online)
DOI 10.58369/biit.v2i3.167
https://ejurnal.amikstiekomsu.ac.id/index.php/BIIT

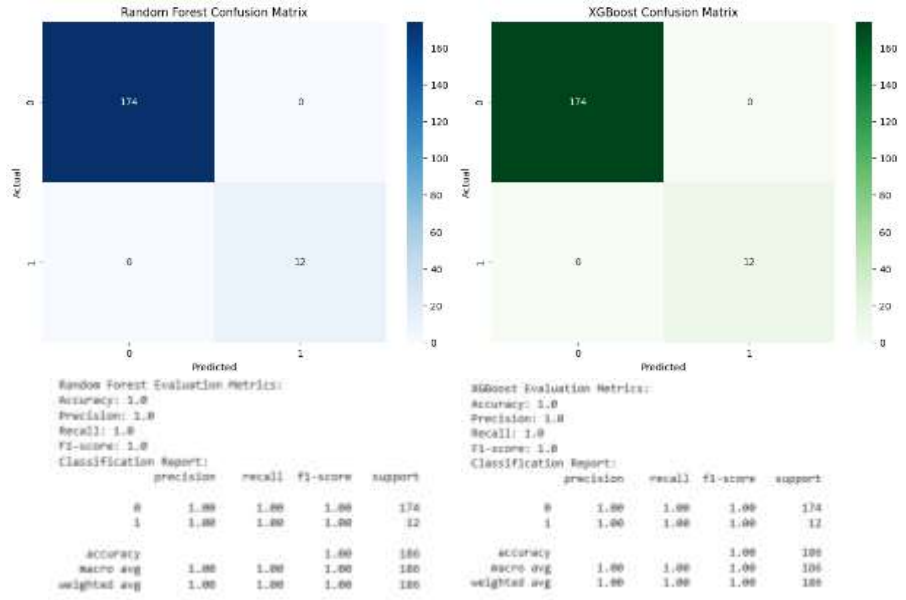


Figure 4 visualization of the Matrix

The models have identical evaluation performance, and the table above shows that the second model has identical and perfect evaluation performance based on the classification dimension. All metrics show a value of 1 based on the classification metric. All metrics show a value of 1.00, which means there are no classification errors in the predictions on the test data.

3.4 Testing the SHapley Additive Explanations Model

Use SHAP values to understand the importance of features and contributions to the best Random Forest and XGBoost models.

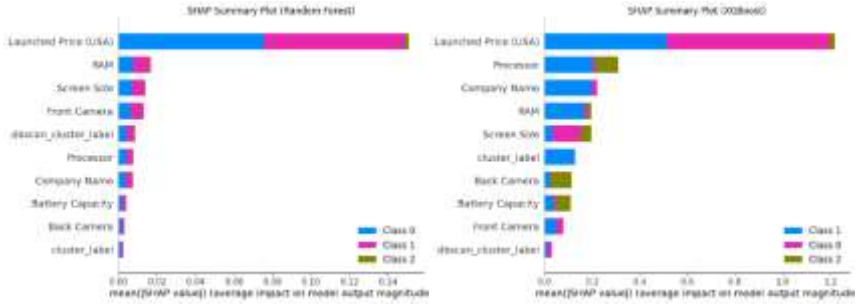


Figure 5. SHapley Additive Explanations Testing

The first figure is a SHAP Summary Plot for the Random Forest model, which illustrates the contribution of each feature to the Random Forest model, showing the contribution of each feature to the prediction model. From the visualization, it can be seen that the "Launched Price (USA)" feature has the most dominant influence on the model output, as indicated by the significantly longer bar compared to other features. This indicates that the Random Forest model is highly dependent on launch price information to classify data into the appropriate class. In addition, features such as RAM, screen size, and front camera also contribute, albeit with much smaller influences. The colors in the graph indicate the contribution to each class, where blue, pink, and olive green represent Class 0, Class 1, and Class 2, respectively.
Meanwhile, the second image is the SHAP Summary Plot for the XGBoost model. Similar to the Random Forest model, the "Launched Price (USA)" feature is also the most influential variable. However, unlike Random Forest, which is highly dependent on one main feature, XGBoost shows a more even contribution. Other features such as Processor, Company Name, RAM, and Screen Size show a significant influence on the prediction results. This indicates that the XGBoost model utilizes more information from various features simultaneously in the decision-making process. In addition, the

Buletin Ilmiah Informatika Teknologi
Vol 4, No 1, September, Hal. 25 - 32
ISSN 2962-0945 (media online)
DOI 10.58369/biit.v2i3.167
https://ejurnal.amikstiekomsu.ac.id/index.php/BIIT

contribution of features to each class appears more balanced, which shows the higher complexity and generalization ability of the XGBoost model compared to Random Forest.
Overall, these two visualizations provide a clear picture of how each model utilizes features. Random Forest tends to focus on dominant single features, while XGBoost adopts a more holistic approach and considers interactions between various input variables. This understanding can be

4. CONCLUSION

In this study, clustering using the DBSCAN algorithm showed significant advantages over the K-Means method, especially in terms of cluster separation. With the eps parameter set at 0.3 and min_samples at 2, DBSCAN achieved a Silhouette Score of 0.927, which was higher than K-Means, which only achieved 0.9069. This indicates that DBSCAN is more effective in identifying clearer and more distinct cluster structures in the analyzed data.
On the other hand, in terms of classification, both the Random Forest and XGBoost models, which have been optimized, show excellent performance with perfect accuracy (1.0) on the test set. Although these results appear promising, further investigation is necessary, as very high accuracy may indicate overfitting. Overfitting occurs when a model is too complex and too well-fitted to the training data, thereby reducing its ability to generalize to new data.
Additionally, feature importance analysis using the SHAP (SHapley Additive exPlanations) method was conducted to provide deeper insights into the contribution of each feature to the model's predictions. However, although summary plots were successfully generated, there were challenges in producing dependency plots and strength plots, which hindered a more comprehensive understanding of the interactions between features. This highlights the need for improvements in the analysis process to ensure a more in-depth and accurate interpretation of the model used..

REFERENCES

[1] A. Kapoor, A. Sahay, N. C. Singh, V. S. Chandrasekhar Pammi, and P. Banerjee, “The neural correlates and the underlying processes of weak brand choices,” *J Bus Res*, vol. 154, p. 113230, Jan. 2023, doi: 10.1016/J.JBUSRES.2022.07.056.

[2] A. Kapoor, A. Sahay, N. C. Singh, V. S. Chandrasekhar Pammi, and P. Banerjee, “The neural correlates and the underlying processes of weak brand choices,” *J Bus Res*, vol. 154, p. 113230, Jan. 2023, doi: 10.1016/J.JBUSRES.2022.07.056.

[3] R. A. Hasan, H. Irshaid, F. Alhomaidat, S. Lee, and J. S. Oh, “Transportation Mode Detection by Using Smartphones and Smartwatches with Machine Learning,” *KSCE Journal of Civil Engineering*, vol. 26, no. 8, pp. 3578–3589, Aug. 2022, doi: 10.1007/S12205-022-1281-0.

[4] A. L. Potzel, C. Gar, J. Seissler, and A. Lechner, “A Smartphone App (TRIANGLE) to Change Cardiometabolic Risk Behaviors in Women Following Gestational Diabetes Mellitus: Intervention Mapping Approach,” *JMIR Mhealth Uhealth*, vol. 9, no. 5, May 2021, doi: 10.2196/26163.

[5] S. Moro, G. Pires, P. Rita, and P. Cortez, “A cross-cultural case study of consumers’ communications about a new technological product,” *J Bus Res*, vol. 121, pp. 438–447, Dec. 2020, doi: 10.1016/J.JBUSRES.2018.08.009.

[6] M. Ueki, “A deflation-adjusted Bayesian information criterion for selecting the number of clusters in K-means clustering,” *Comput Stat Data Anal*, vol. 209, p. 108170, Sep. 2025, doi: 10.1016/J.CSDA.2025.108170.

[7] J. Lu, T. Luo, and K. Li, “A forward k-means algorithm for regression clustering,” *Inf Sci (N Y)*, vol. 711, p. 122105, Sep. 2025, doi: 10.1016/J.INS.2025.122105.

[8] Y. Dai, L. Yang, and Y. Cao, “Long baseline underwater source localization based on deep K-Means++ clustering in complex underwater environments,” *Digit Signal Process*, vol. 164, p. 105281, Sep. 2025, doi: 10.1016/J.DSP.2025.105281.

[9] J. Li, L. Wang, S. Fu, W. Fu, and X. Pan, "Self-labeled framework with semi-supervised ball K-means clustering-based synthetic example generation for semi-supervised classification in industrial applications," *Eng Appl Artif Intell*, vol. 150, p. 110528, Jun. 2025, doi: 10.1016/J.ENGAPPAI.2025.110528.

[10] A. A. Hussein, H. N. Abdulrazzak, and A. S. Ali, "MANET highly efficient clustering technique based on coverage k-means algorithm," *Egyptian Informatics Journal*, vol. 30, p. 100672, Jun. 2025, doi: 10.1016/J.EIJ.2025.100672.

[11] P. Bíró, B. B. Bálint, T. Novák, and M. Erdélyi, "Cluster parameter-based DBSCAN maps for image characterization," *Comput Struct Biotechnol J*, vol. 27, pp. 920–927, Jan. 2025, doi: 10.1016/J.CSBJ.2025.02.037.

[12] R. Ma, J. Sha, S. Zhang, D. Zhu, W. Kang, and J. Liu, "Fast grouping fusion method of dual carbon monitoring data based on DBSCAN clustering algorithm," *Results in Engineering*, vol. 26, p. 105057, Jun. 2025, doi: 10.1016/J.RINENG.2025.105057.

[13] Z. Wu, X. Fan, G. Bian, Y. Liu, X. Zhang, and Y. Q. Chen, "Short-term wind power forecast with turning weather based on DBSCAN-RFE-LightGBM," *Renew Energy*, vol. 251, p. 123217, Oct. 2025, doi: 10.1016/J.RENENE.2025.123217.

[14] P. Zhang, J. Duan, C. Wang, X. Li, J. Su, and Q. Shang, "Predicting response to anti-VEGF therapy in neovascular age-related macular degeneration using random forest and SHAP algorithms," *Photodiagnosis Photodyn Ther*, p. 104635, May 2025, doi: 10.1016/J.PDPDT.2025.104635.

[15] H. Zhu, G. Liu, X. Gao, S. Wang, and C. Jiang, "A Random Forest-Based Combinatorial Optimization Model for Altitude-Dependent Vertical Correction of Precipitable Water Vapor in China," *Advances in Space Research*, May 2025, doi: 10.1016/J.ASR.2025.05.039.

[16] X. Wei *et al.*, "Study on Prediction Model of Nitrogen Oxide Concentration in Reprocessing Plant Based on Random Forest," *International Journal of Advanced Nuclear Reactor Design and Technology*, May 2025, doi: 10.1016/J.JANDT.2025.04.011.

[17] H. Liu *et al.*, "Research on time-domain motion prediction of floating platforms based on XGBoost model," *Ocean Engineering*, vol. 332, p. 121393, Jul. 2025, doi: 10.1016/J.OCEANENG.2025.121393.

[18] S. Zhou, "Gwo-ga-xgboost-based model for Radio-Frequency power amplifier under different temperatures," *Expert Syst Appl*, vol. 278, p. 127439, Jun. 2025, doi: 10.1016/J.ESWA.2025.127439.

[19] X. Zhao, P.-F. Zhang, Q. Zhao, D. Zhang, Y. Tuerxunmaimaiti, and H. Cao, "A SHAP algorithm-based prediction of the interlaminar shear strength degradation of G/BFRP bars embedded in concrete exposed to marine environment," *Case Studies in Construction Materials*, p. e04770, May 2025, doi: 10.1016/J.CSCM.2025.E04770.

[20] Y. Sun and H. Ma, "Interpretable analysis of transformer winding vibration characteristics: SHAP and multi-classification feature optimization," *International Journal of Electrical Power & Energy Systems*, vol. 166, p. 110585, May 2025, doi: 10.1016/J.IJEPES.2025.110585.

[21] M. Kruk, "SHAP-NET, a network based on Shapley values as a new tool to improve the explainability of the XGBoost-SHAP model for the problem of water quality," *Environmental Modelling & Software*, vol. 188, p. 106403, Apr. 2025, doi: 10.1016/J.ENVSOFT.2025.106403.

Buletin Ilmiah Informatika Teknologi
Vol 4, No 1, September, Hal. 25 - 32
ISSN 2962-0945 (media online)
DOI 10.58369/biit.v2i3.167
https://ejurnal.amikstiekomsu.ac.id/index.php/BIIT

[22] R. Hermawati and I. S. Sitanggang, "Web-Based Clustering Application Using Shiny Framework and DBSCAN Algorithm for Hotspots Data in Peatland in Sumatra," *Procedia Environ Sci*, vol. 33, pp. 317–323, Jan. 2016, doi: 10.1016/J.PROENV.2016.03.082.

[23] A. Arafa, N. El-Fishawy, M. Badawy, and M. Radad, "RN-SMOTE: Reduced Noise SMOTE based on DBSCAN for enhancing imbalanced data classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 5059–5074, Sep. 2022, doi: 10.1016/J.JKSUCI.2022.06.005.

[24] K. S. Pranata, A. A. S. Gunawan, and F. L. Gaol, "Development clustering system IDX company with k-means algorithm and DBSCAN based on fundamental indicator and ESG," *Procedia Comput Sci*, vol. 216, pp. 319–327, Jan. 2023, doi: 10.1016/J.PROCS.2022.12.142.

[25] N.-T. Ho *et al.*, "MACHINE LEARNING APPROACH WITH RANDOM FOREST AND SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE MAY BE OPTIMAL IN NON-INVASIVE EUPLOIDY DETECTION: A PRELIMINARY STUDY," *Fertil Steril*, vol. 120, no. 4, p. e107, Oct. 2023, doi: 10.1016/J.FERTNSTERT.2023.08.352.

[26] L. Xu, S. Wen, H. Huang, Y. Tang, Y. Wang, and C. Pan, "Corrosion failure prediction in natural gas pipelines using an interpretable XGBoost model: Insights and applications," *Energy*, vol. 325, p. 136157, Jun. 2025, doi: 10.1016/J.ENERGY.2025.136157.

[27] F. Zhang, Z. Zhu, J. Liu, Y. Zhang, M. Xu, and P. Jia, "A novel concentration prediction technique of carbon monoxide (CO) based on beluga whale optimization-extreme gradient boosting (BWO-XGBoost)," *J Taiwan Inst Chem Eng*, vol. 171, p. 106045, Jun. 2025, doi: 10.1016/J.JTICE.2025.106045.

[28] T. K. Yu, I. C. Chang, S. De Chen, H. L. Chen, and T. Y. Yu, "Predicting potential soil and groundwater contamination risks from gas stations using three machine learning models (XGBoost, LightGBM, and Random Forest)," *Process Safety and Environmental Protection*, vol. 199, p. 107249, Jul. 2025, doi: 10.1016/J.PSEP.2025.107249.