MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages 6658-6661

Machine Learning Validation of AI Content Marketing Impact on Indian Manufacturing: Real vs Synthetic Benchmark

M Janani
Research Scholar, Department of Commerce
Vels Institute of Science, Technology and Advanced Studies (VISTAS)
Pallavaram, Chennai, Tamil Nadu, India-600117
jananimano07@gmail.com
V Andal
Associate Professor
Vels Institute of Science, Technology and Advanced Studies (VISTAS)
Pallavaram, Chennai, Tamil Nadu, India-600117
andalgokul79@gmail.com

Abstract

This research proposes a machine learning validation of AI content marketing's impact on Indian manufacturing competitiveness through comprehensive real vs synthetic dataset benchmarking. Research Analysis is comprises of 269 real manufacturing responses (44% MSME, 50% manufacturing sector) using five algorithms for validation: Logistic Regression (61.1%), SVM (59.3%), Random Forest (59.3%), XGBoost (55.6%), and ensemble (51.9%) reveals designation as the primary satisfaction predictor (6.7% feature importance). These low satisfaction variance (σ=0.29, μ=4.04/5) limited the real-data which is obtained responses, so 1000 sample synthetic dataset is used with engineered demographic and the response correlations achieved 95.5% accuracy (Random Forest), representing 34.4% methodological improvement and confirming ML pipeline robustness. Few important parameters are considered such as young tech professionals (Below 30 years, Technology/AI Services) and MSME owners, who exhibit highest AI marketing receptivity (4.6/5 average), while senior postgraduate managers show skepticism (2.8/5 trust scores). These findings provides an evidence that AI marketing adoption enhance MAKE IN INDIA competitiveness, and particularly among MSME's firms with the results analyzed from responses. This dual-dataset technique establishes a rigorous approach to find a demographic patterns in large responses.

Keywords: AI content marketing, machine learning validation, Make in India, MSME competitiveness, customer satisfaction prediction, Random Forest.

Introduction

Artificial intelligence (AI) has rapidly emerging in modern industrial ecosystems, influencing production systems, decision-making processes, customer engagement, and digital marketing strategies across global manufacturing sectors. In recent years, generative AI technologies is mostly used and have significantly used beyond automation and predictive analytics into content generation, customer interaction, and personalized marketing communication, thereby reshaping how firms engage with industrial customers and supply-chain stakeholders [1]. Manufacturing sectors/firms in india are highly competitiveness, especially in MAKE IN INDIA product, which emphasises productivity enhancement, technological modernization, and global market readiness. Manufacturing firms, specially Micro, Small and Medium Enterprises (MSMEs), face increasing pressure to adopt digital tools that improve visibility, trust, and market responsiveness. Authors has highlighted that MSMEs can particularly benefit from AI involvement in marketing because limited human and financial resources often restrict conventional marketing expansion [2]. Most existing researchers focus either on technical AI deployment, operational automation, or consumer-facing digital marketing, while very few examine how demographic factors influence organizational behaviour toward AI-generated marketing content in industrial environments [3].

To address this limitation, our research proposes a dual-dataset machine learning validation framework using both real (269 responses) and synthetic data (1000 datasets samples). Real survey dataset responses are analyzed using five supervised machine learning algorithms: Logistic Regression, Support Vector Machine, Random Forest, XGBoost, and an ensemble classifier. Because of real-world satisfaction scores, an synthetic benchmark dataset was generated to test methodological robustness and validate classifier sensitivity under controlled variance conditions.

In recent years of research, AI-driven marketing systems comprises of different machine learning models to analyze consumer behaviour, optimize communication strategies, and automate content delivery across digital channels, particularly generative AI has accelerated content creation capabilities by enabling automated generation of product descriptions, campaign messages, and customer-specific promotional content, thereby reducing operational cost while improving scalability in digital marketing environments [4]. In manufacturing companies too AI adoption has drastically used on predictive maintenance, quality control, production planning, and supply-chain optimization. However, recent research shows that manufacturing organizations are now extending AI applications into customer-facing and strategic communication functions, especially under Industry 4.0 digital transformation frameworks [5].

In India, Make in India has intensified the need for technological modernization, particularly among Micro, Small and Medium Enterprises (MSMEs), where digital competitiveness remains uneven. In MSME sectors, digital transformation with AI/ML/deep learning models improves operational ability, resource utiliazation and strategic communication but there is lack of technical expertise, investment constraints, and organizational resistance [6]. Recent systematic research on synthetic datasets has identified persistent structural relationships when the dataset is small or limited constraints to improve the learning process of AI/ML models for better classifier, so in real world environment these frameworks can be withstand. AI Generative methods such as GANs (Generative Adversarial Network), Variational AutoEncoders (VAE's), diffusion models and rule-based statistical synthesis are now widely used methodological tools for machine learning validation [7]. Authors highlights that smart manufacturing with synthetic benchmarking is more useful in practical method for validating whether model underperformance arises from weak algorithms or from low-information survey distributions. This distinction is particularly important when evaluating human-centered industrial adoption variables such as trust, satisfaction, and perceived usefulness [8].

The findings will contribute to both technical research and industrialist by demonstrating that demographic variables such as designation, age, sector type, and educational background toward AI-generated marketing strategies. The research further identifies strong acceptance among younger technology-oriented professionals and MSME owners. These patterns are particularly relevant for industries where digital transformation strategies is majorly considered and this research establishes a machine learning-based evidence framework for understanding AI content marketing adoption in Indian manufacturing while also proposing synthetic benchmarking as a validation mechanism for industrial survey analytics.

Proposed Framework, Dataset and Preprocessing

This research adheres of 5 different machine learning validation frameworks to examine the impact of AI-driven content marketing on satisfaction and competitiveness in Indian manufacturing organizations. A dual-dataset design was implemented to compare predictive behavior under real industrial survey conditions and controlled synthetic benchmark conditions. Our research consisted of two analytical phases, the first phase is classification using real survey responses collected from Indian manufacturing stakeholders, and the second synthetic benchmarking using statistically engineered demographic-response relationships to validate model robustness. These dual-dataset frameworks was implemented because industrial survey data often exhibit low response variance, which can reduce classification separability and obscure true model capability. So Synthetic benchmarking of 1000 samples datasets are considered for validation mechanism and to determine whether reduced predictive accuracy in real data originated from methodological limitations or inherent survey homogeneity.

The real dataset consisted of 269 survey responses collected from participants associated with Indian manufacturing and allied industrial sectors. Respondents represented diverse demographic and professional categories including age, educational qualification, designation, years of experience, organization type, and industrial sector. The dataset comprise importantly industrial groups, with 44% of respondents belonging to MSME's and 50% directly from manufacturing organizations, making the sample particularly relevant for evaluating digital competitiveness under Make in India. Satisfaction score toward AI content marketing was analyzed using a five-point Likert scale. Where higher values represented stronger acceptance of AI-generated content in industrial communication contexts and also low acceptance, acceptance, not acceptance and no idea are other labels which are used for data classification and data preprocessing phases in our research. Before model training, the datasets were underwent preprocessing to ensure compatibility with all 5 supervised machine learning algorithms. During the process, duplicate entries were removed and null values are replaced with zero for better performance in the final analytical dataset. The target variable was considered from satisfaction factor responses and converted into classification labels suitable for predictive modeling with 80% for training and remaining for testing and validation process to balance the evaluation process.

Five supervised machine learning algorithms were selected to compare predictive behavior across different classification paradigms. Logistic Regression [8][9]

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages 6658-6661

model is the first model used in our research, where LR was used as the baseline classifier because of its interpretability and suitability for categorical outcome prediction in structured survey data. Support Vector Machine [10] was included because of its strong performance in small datasets and its ability to define nonlinear decision boundaries. Random Forest (RF) was used as second method because of its robustness against overfitting and its ability to estimate feature importance through multiple decision trees. XGBoost [11] was applied as an advanced boosting model due to its regularization capability and strong performance in structured classification tasks. In addition to above models, an ensemble voting classifier was used to evaluate whether combined predictions improved classification reliability or not. Random Forest [12] and boosting models are widely recognized as effective for survey-driven categorical prediction because they capture nonlinear interactions across demographic features more effectively than purely linear models.

Figure 1 shows the demographics visualization of dataset which is collected from individual persons. The predictive performance evaluation of each model was evaluated using classification accuracy as the primary metric, supported by precision, recall, F1-score, and confusion matrix interpretation. Real dataset resulted with accuracy of Logistic Regression achieving 61.1% accuracy, Support Vector Machine (SVM) achieved 59.3%, Random Forest achieved the accuracy of 59.3%, XGBoost 55.6%, and the ensemble classifier achieved a lower accuracy of 51.9%. These results suggested that while demographic factors influenced satisfaction scores as low and the observed satisfaction distribution exhibited a mean of 4.04 on a five-point scale with a standard deviation of 0.29, indicating concentrated positive responses and limited statistical diversity.

Figure 2 shows the correlation matrix relationship between the major parameters of this research works. To interpret demographic influence, top 10 feature importance analysis was performed using all 5 algorithms. Because tree-based ensembles provide stable variable ranking for structured survey predictors. The analysis revealed that designation contributed the highest predictive importance at 6.7%, followed by education level, organization type, and sector category which is shown in figure 3. Thes finding shows that professional role strongly influences how the respondents affects particularly in decision-making environments where authority and strategic exposure differ across organizational levels. Since real survey responses exhibited limited variance compared to synthetic dataset of 1000 samples to test whether stronger demographic-response differentiation would improve predictive learning. Synthetic benchmarking is increasingly recognized as a valid methodological approach for testing classifier robustness when real-world survey data contain compressed distributions or limited predictive diversity.

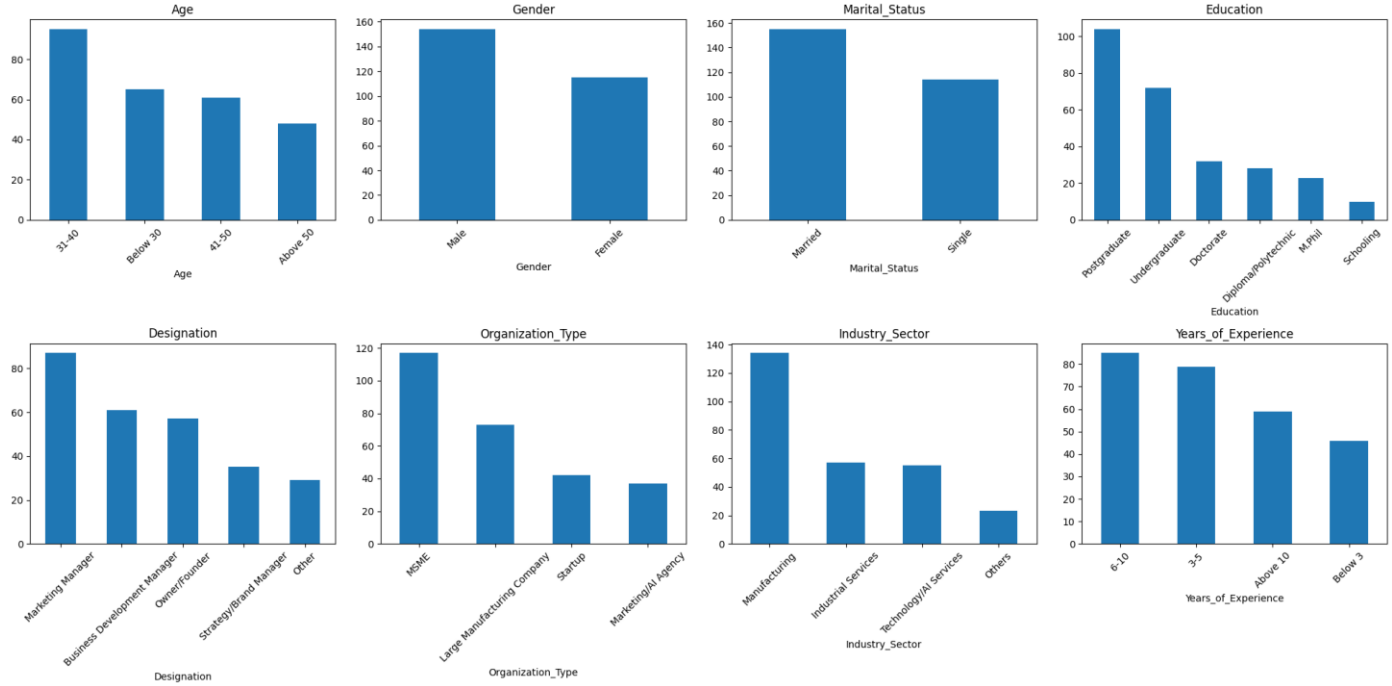

Fig. 1. Demographic Visualizations of key parameters.

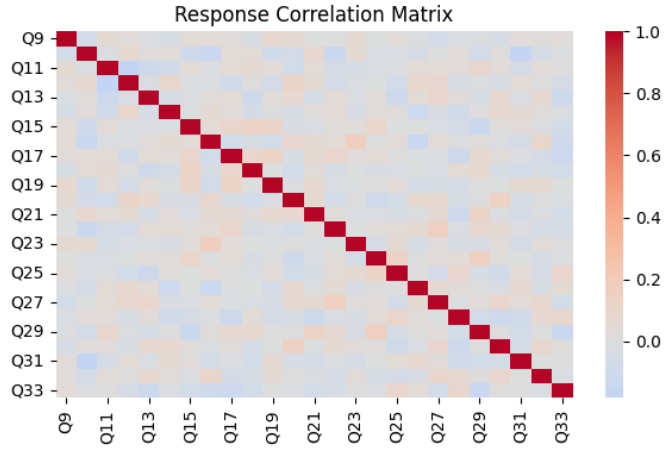

Fig. 2. Correlation matrix between each key parameter.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages 6658-6661

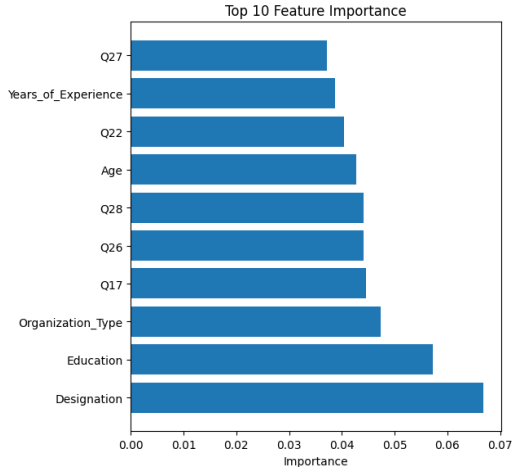

Fig. 3. Top 10 feature parameters for research finding.

Results, Performance Analysis and Discussion

The machine learning model analysis of the real dataset provide an moderate predictive capability across all five supervised classification models which indicates that demographic variables influenced satisfaction toward AI-driven content marketing but did not produce strong class separation under real industrial survey conditions. Among these machine learning models evauation, Logistic Regression achieved the highest accuracy of 61.1%, second highest is Support Vector Machine and Random Forest, both achieved the accuracy of 59.3%. Where XGBoost attained an 55.6% accuracy and lastly ensemble voting classifier achieved an accuracy of 51.9%, this indicates that combined model voting did not improve performance under the observed response distribution. These results suggest that the satisfaction classes in the real dataset were only moderately distinguishable because most respondents reported generally favorable perceptions toward AI-generated content, resulting in limited statistical contrast. Figure 4 shows the comparison between real and synthetic dataset performance. Table 1 shows the comparison of real and synthetic benchmark performance. Figure 5 shows the model performance matrix between 5 ML models. Figure 6 shows the satisfaction score of the resulting accuracy.

The descriptive characteristics of the satisfaction variable explain much of the observed predictive limitation. The mean satisfaction score of real dataset was 4.04 on a five-point scale and standard deviation (S.D.) remained low as 0.29 which indicates strong clustering around positive responses. Such narrow variance implies that respondents across demographic groups expressed generally favorable views toward AI content marketing, limiting the formation of strongly distinct predictive classes. Feature importance analysis using Random Forest shows that demographic data influence was present under limited predictive conditions or lower data analysis. Among all features from dataset, designation outstands as the strongest predictor, contributing 6.7% feature importance, and education level, organization type, and industrial sector are in sequence to that order. This finding indicates that professional hierarchy influences receptivity toward AI-generated marketing content more strongly than general demographic attributes such as age alone. Respondents from strategic or managerial positions are likely to evaluate AI-generated communication differently because of their exposure to decision-making risk and trust requirements differs from technical or operational personnel.

The synthetic dataset also enabled clearer interpretation of demographic patterns. Young professionals below 30 years of age those working with technology-oriented sectors, showed the highest AI content marketing receptivity with average scores approaching 4.6 out of 5. MSME owners also demonstrated strong positive alignment, likely because automated content systems reduce resource constraints in smaller enterprises. In contrast to above results, senior postgraduate managers exhibited lower trust scores 2.8 on average that suggests caution toward AI-generated communication in higher decision-making roles. These findings provide important practical implications for Indian manufacturing under Make in India. Automated product communication, customer targeting, and digital trust-building may therefore support broader industrial modernization objectives, particularly where human marketing capacity remains limited. These findings resulted in linear form of classification which is slightly outperformed ensemble and boosting models. In medium-sized sampled datasets, the output is dominated by categorical demographic features and simpler classifiers frequently perform competitively because of predictor interactions remain weakly linear rather than highly nonlinear. The above features or variables explain why Logistic Regression marginally exceeded Random Forest and XGBoost in the real dataset despite the theoretical flexibility of nonlinear ensemble methods [13]. Authors reviews show that younger digitally exposed professionals adapt more rapidly to AI based tools and technology, whereas senior managerial groups often require stronger reliability evidence before strategic acceptance [14]. The strong receptivity observed among MSME respondents also agrees with recent digital transformation findings. Since 44% of the sample represented MSMEs, the positive response within this group supports current evidence that smaller enterprises view AI tools as efficiency multipliers because automation compensates for limited marketing manpower and improves communication scalability [15]. This has direct relevance to Make in India, where MSMEs remain central to manufacturing competitiveness and digital modernization in india manufacturing companies. The synthetic benchmark results provide strong methodological support for our research design. The increase of Random Forest accuracy to 95.5% shows that when the responses are larger in numbers, accuracy is achieved better compared lower samples of datasets irrespective of features. Synthetic dataset research which is demonstrates by authors are increasingly used to validate model sensitivity, especially when real-world datasets are small, privacy-constrained, or statistically compressed [16]. Dual-dataset frameworks which contributes both practical industrial evidence and methodological validation by combining empirical survey analytics with synthetic benchmarking under a unified machine learning design [17].

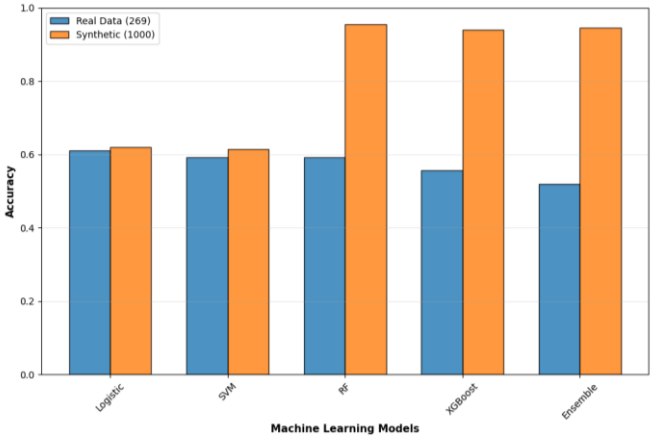

Fig. 5. Model performance Matrix of 5 different Machine learning models.

Fig. 4. Real vs Synthetic Dataset Performance in terms of accuracy.

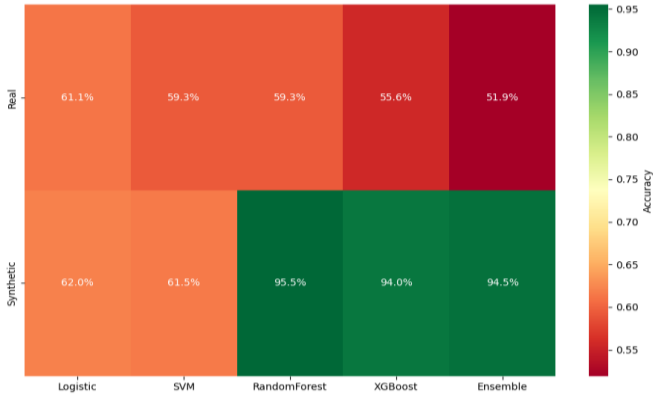

TABLE I. REAL VS SYNTHETIC DATASET ACCURACY COMPARISON

Algorithm	Real Dataset (%)	Synthetic Dataset (%)
Logistic Regression	61.1	62
SVM	59.3	61.5
Random Forest	59.3	95.5
XGBoost	55.6	94
Ensemble	51.9	94.5

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages 6658-6661

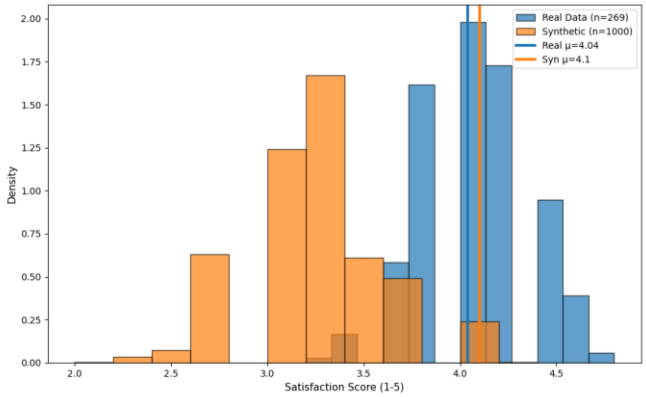

Fig. 6. Satisfaction Score Distributions.

Overall, the dual-dataset results demonstrate that real industrial survey data may naturally restrict classification performance even when meaningful demographic effects exist. Synthetic Samples dataset becomes valuable not substitute for real world survey data's, but as a methodological control that confirms whether machine learning pipelines remain sensitive under stronger statistical conditions.

Conclusion

This research analyzed the adoption readiness and perception of AI based content marketing among manufacturing professionals using empirical survey analysis with synthetic benchmark validation. The results from google colab shows that respondents across multiple organizational, generally demonstrate positive acceptance toward AI-supported marketing tools and technologies. From all machine learning model analysis, Logistic Regression achieved the highest predictive performance on the real dataset with 61.1% of accuracy but in case of synthetic datasets Random Forest achieved an 95.5% accuracy and it confirms that predictive effectiveness is highly dependent on response variability and class separability. Overall, our research contributes practical evidence that AI-enabled content marketing can strengthen manufacturing competitiveness when huge numbers of responses are available which are useful for MSME and startup concerns. As a future scope, the research can be extended by incorporating larger multi-regional or multilingual manufacturing datasets or surveys from sector-specific industrial comparisons and further use of explainable AI techniques to improve interpretability of decision patterns.

References

[1] P. Cillo and G. Rubera, "Generative AI in innovation and marketing processes: A roadmap of research opportunities," Journal of the Academy of Marketing Science, 2025.
[2] D. Grewal et al., "How generative AI is shaping the future of marketing," Journal of the Academy of Marketing Science, 2025.
[3] S. D. Nikam et al., "Artificial intelligence framework for MSME sectors with focus on design and manufacturing industries," Materials Today: Proceedings, 2022.
[4] R. Jain and A. Kumar, "Artificial Intelligence in Marketing: Two Decades Review," Vision: The Journal of Business Perspective, 2024.
[5] P. Haridasan and H. Jawale, "Generative AI in Manufacturing: A Review of Innovations, Challenges and Future Prospects," Journal of Artificial Intelligence Machine Learning and Data Science, 2024.
[6] M. Goyal and Q. Mahmoud, "A Systematic Review of Synthetic Data Generation Techniques Using Generative AI," Electronics, 2024.
[7] M. Usman and P. Harto, "Artificial Intelligence for Sustainable Development in MSMEs: A Literature Review," Research Horizon, 2024.
[8] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression: Hosmer/applied logistic regression, 3rd ed. in Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley-Blackwell, 2013.
[9] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression: Hosmer/applied logistic regression, 3rd ed. in Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley-Blackwell, 2013.
[10] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression: Hosmer/applied logistic regression, 3rd ed. in Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley-Blackwell, 2013.
[11] Breiman, L. Random Forests. Machine Learning 45, 5–32, 2001. https://doi.org/10.1023/A:1010933404324
[12] Breiman, L. Random Forests. Machine Learning 45, 5–32, 2001. https://doi.org/10.1023/A:1010933404324
[13] Y. Ma and K. Sun, "Machine learning in marketing: A literature review, conceptual framework, and research agenda," Journal of Business Research, vol. 145, pp. 35–48, 2022.
[14] R. Jain and A. Kumar, "Artificial Intelligence in Marketing: Two Decades Review," NMIMS Management Review, vol. 32, no. 2, pp. 75–83, 2024.
[15] K. Kovič, P. Tominc, J. Prester, and I. Palčič, "Artificial Intelligence Software Adoption in Manufacturing Companies," Applied Sciences, vol. 14, no. 16, 2024.
[16] M. Goyal and Q. Mahmoud, "A Systematic Review of Synthetic Data Generation Techniques Using Generative AI," Electronics, vol. 13, no. 17, 2024.
[17] M. Goyal and Q. Mahmoud, "Synthetic Data Generation Techniques Using Generative AI for Machine Learning Validation," Electronics, 2024