Development of a Sales Prediction Model Using Support Vector Machine to Support Digital Transformation in MSMEs

Eko Purwanto[1], Bangun Prajadi Cipto Utomo[2], Hanifah Permatasari[3], Farahwahida Mohd[4]
[1,3]Information System, Faculty of Computer Science, Universitas Duta Bangsa, Surakarta, Indonesia
[2]Management, Faculty of Law and Business, Universitas Duta Bangsa, Surakarta, Indonesia
[4] Malaysian Institute Information Technology, Universiti Kuala Lumpur, Malaysia
[1]eko_purwanto@udb.ac.id , [2]bangun_prajadi@udb.ac.id, [3]hanifah_permatasari@udb.ac.id, [4]farawahidam@unikl.edu.my

Abstract— The digital transformation of Micro, Small, and Medium Enterprises (MSMEs) is crucial for enhancing operational efficiency and competitiveness in an increasingly digital world. One key aspect of digitalization is the ability to accurately predict sales, which facilitates better decision-making and resource planning. This study aims to develop a predictive sales model using Support Vector Machine (SVM), optimized through GridSearchCV to find the best combination of hyperparameters. The dataset includes variables such as price, stock, and promotion as features, while sales is the target variable. The optimal combination of hyperparameters identified includes C = 0.1, gamma = 0.1, and kernel = 'poly', providing a robust model for sales prediction. The model yielded a Mean Squared Error (MSE) of 2105.94 and a R^2 (Coefficient of Determination) of -0.0293, indicating that the model has strong predictive capabilities despite its potential for further optimization. This model serves as a foundation for sales forecasting in MSMEs, offering valuable insights into how artificial intelligence can support efficient decision-making. The results highlight the potential for future improvements in model accuracy, enabling MSMEs to better leverage data-driven strategies to enhance business performance and achieve sustainable growth in the digital age.

Keywords — Development; Sales Prediction; Support Vector Machine; Digital Transformation

I. INTRODUCTION

The digital transformation of Micro, Small, and Medium Enterprises (MSMEs) has become essential for improving operational efficiency and competitiveness in today's rapidly evolving market [1][2][3]. As businesses increasingly rely on data-driven strategies, sales prediction has emerged as a critical area for improving decision-making in inventory management, pricing, and resource allocation [4]. Despite the advancements, many MSMEs still face challenges in leveraging predictive analytics due to limited computational resources and analytical expertise [5][6].

Previous studies have explored various techniques for sales forecasting, including statistical models and machine learning algorithms. Machine learning, particularly Support Vector Machine (SVM), has gained significant attention due to its ability to model complex, non-linear relationships in data [7][8]. Researchers like [9] and [10] have demonstrated the effectiveness of SVM in retail and e-commerce applications for predicting demand and sales. However, few studies have applied SVM for sales prediction in MSMEs, especially considering the unique challenges they face in the digital transformation journey.

This research aims to fill this gap by developing a sales prediction model using SVM, optimized through GridSearchCV to identify the best hyperparameters for improved prediction accuracy [11][12]. The model incorporates key business features such as price, stock, and promotion, with sales as the target variable. By fine-tuning parameters like C, gamma, and kernel, this study seeks to provide a robust solution for MSMEs looking to enhance their sales forecasting capabilities [13][14].

The findings of this research are expected to provide practical insights for MSMEs, enabling them to leverage predictive analytics for better decision-making and more efficient resource allocation [6][15]. The results will also contribute to the growing body of knowledge on machine learning applications in MSMEs' digital transformation.

II. METHOD

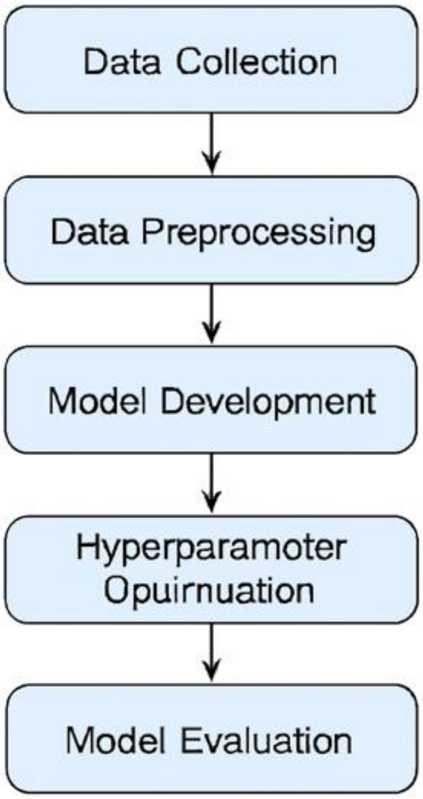

Figure 1. Research Methodology

A. Data Collection

This study utilizes historical sales data from Micro, Small, and Medium Enterprises (MSMEs) to build a predictive sales model using Support Vector Machine (SVM). The data includes three predictor variables: price, stock, and promotion, with sales as the target variable [5]. The dataset consists of 500 entries representing daily transaction data over the past three years. This data was sourced from the internal records of a selected MSME and is representative of typical transactional data in a retail business.

B. Data Preprocessing

To ensure the quality and consistency of the data, several preprocessing steps were carried out before using the data to train the model:

1. Data Cleaning
   The dataset was first cleaned to remove any missing or duplicate entries. Missing values in numerical columns were imputed using the mean imputation method.
2. Normalization
   All numerical features (price, stock, and promotion) were normalized using StandardScaler to ensure they have a mean of 0 and a standard deviation of 1. This is crucial for models like SVM, which are sensitive to the scale of the data.
3. Data Splitting
   The dataset was randomly split into two subsets: 80% for training and 20% for testing. The training set was used to build the model, while the testing set was used to evaluate the model's predictive accuracy.

C. Model Development

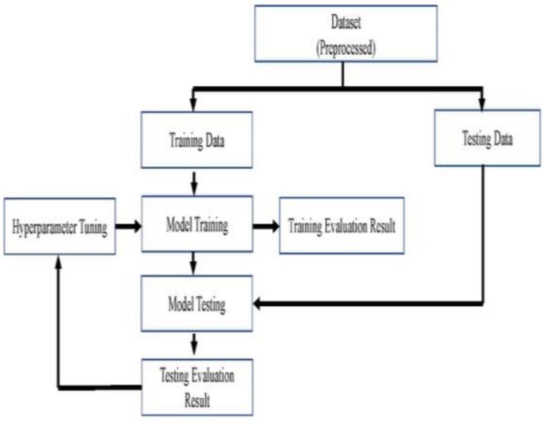

Figure 2. Model Sales Prediction

The primary model used in this study is Support Vector Machine (SVM), specifically the SVM Regressor for predicting continuous values such as sales. The SVM algorithm was selected due to its capability to handle high-dimensional data and model complex, non-linear relationships [16]. The SVM aims to find the optimal hyperplane that best separates the data, and in the case of regression, it predicts continuous values such as sales.

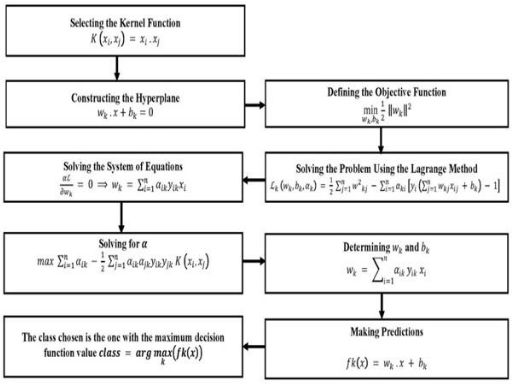

Figure 3. Sales Prediction Use Support Vector Machine

Step 1: Selecting the Kernel Function

The kernel function chosen for this study is a simple linear kernel, defined by the basic equation (1):

$$K\left(x_i, x_j\right) = x_i . x_j \quad (1)$$

The kernel function computes the dot product in the feature space without explicitly mapping the data to that space. Kernels help in solving non-linear problems by transforming them into linear problems in higher-dimensional space. The linear kernel is one of the simplest kernel functions, used in SVM to separate data that can be linearly separated.

In a linear kernel, $K\left(x_i, x_j\right)$ is simply the dot product between two feature vectors x_i *and* x_j. If x_i *and* x_j are two vectors in feature space, their dot product is computed as:

$$x_i . x_j = \sum_{k=1}^{n} x_{i,k} \cdot x_{j,k}$$

where $x_{i,k}$ and $x_{j,k}$ are the elements of vectors x_i and x_j

Step 2: Constructing the Hyperplane

For each class k, a hyperplane equation separates that class from others. This hyperplane equation is expressed as:

$$w_k . x + b_k = 0 \quad (2)$$

where

w_k is the weight vector for class k

b_k is the bias for class k

Data is classified according to the following condition:

$$\begin{cases} w_k . x + b_k \geq 1 \, for \, y_i = k \\ w_k . x + b_k \leq -1 \, for \, y_i \neq k \end{cases} \quad (3)$$

Step 3: Defining the Objective Function

To maximize the margin, the objective is to minimize $\|w_k\|$ which is equivalent to minimizing $\frac{1}{2} \|w_k\|^2$ for each class k. Thus, the optimization problem for each class k can be written as:

$$\min_{w_k, b_k} \frac{1}{2} \|w_k\|^2 \quad (4)$$

Step 4: Solving the Problem Using the Lagrange Method

To solve this problem, the Lagrange method is used. The Lagrangian function for class k is given by:

$$\mathcal{L}_k (w_k, b_k, \alpha_k) = \frac{1}{2} \sum_{j=1}^{n} w^2{}_{kj} - \sum_{i=1}^{n} \alpha_{ki} \left[y_i \left(\sum_{j=1}^{n} w_{kj} x_{ij} + b_k \right) - 1 \right] \quad (5)$$

where $\alpha_{ki}$is the Lagrange multiplier, which must be positive $(\alpha_{ki} \geq 0)$

Step 5: Solving the System of Equations

For each k :

$$\frac{\alpha \mathcal{L}}{\partial w_k} = 0 \Rightarrow w_k = \sum_{i=1}^{n} \alpha_{ik} y_{ik} x_i \quad (6)$$

More specifically:

$w_{kj} = \sum_{i=1}^{n} \alpha_{ik} y_{ik} x_{ij}$ for each j

ISSN 2442-7888 (online) DOI 10.24167/sisforma.v12i2.14282

and:

$$\frac{\alpha \mathcal{L}}{\partial b_k} = 0 \Longrightarrow \sum_{i=1}^{n} \alpha_{ik} y_{ik} = 0 \quad (7)$$

$$\alpha_{ik} \left[y_i \left(w_k \, . \, x_i + b_k \right) - 1 \right] = 0$$

Step 6: Solving for α

Using the dual form of the Lagrangian function for each class k:

$$max \sum_{i=1}^{n} \alpha_{ik} - \frac{1}{2} \sum_{j=1}^{n} \alpha_{ik} \alpha_{jk} y_{ik} y_{jk} K\left(x_i, x_j\right) \quad (8)$$

Subject to:

$$\sum_{i=1}^{n} \alpha_{ik} y_{ik} = 0 \quad (9)$$

$$\alpha_{ik} \geq 0$$

Step 7: Determining w_k and b_k

After finding the values of α_{ik}, for each k, w_k and b_k can be determined as:

$$w_k = \sum_{i=1}^{n} \alpha_{ik} \, y_{ik} \, x_i \quad (10)$$

Or more specifically:

$$w_{kj} = \sum_{i=1}^{n} \alpha_{ik} \, y_{ik} \, x_{ij} \ \text{ for each } j \quad (11)$$

For b_k, by selecting any data x_i that satisfies $0 < \alpha_{ik} < C$ we have:

$$b_k = y_{ik} - \sum_{i=1}^{n} \alpha_{ik} y_{ik} \, \mathrm{K}\left(x_i, x_j\right) \quad (12)$$

Step 8: Making Predictions

To classify a new point x using all the models trained for each class, the decision function value for each class k with w_k *and* b_k is given by:

$$fk(x) = w_k \, .x + b_k \quad (13)$$

The class chosen is the one with the maximum decision function value:

$$class = arg \max_k \left(fk(x)\right) \quad (14).$$

III. RESULTS AND DISCUSSION

A. RESULT

The predictive sales model was developed using Support Vector Machine (SVM), with hyperparameter optimization performed using GridSearchCV. The dataset included features such as price, stock, and promotion, with sales as the target variable. After preprocessing and training the model, the best parameters identified by GridSearchCV were C = 0.1, gamma = 0.1, and kernel = 'poly'.

Mean Squared Error (MSE) for the optimized model was 2105.94, while the baseline model (which predicts the mean value of sales) had an MSE of 2500. R^2 (Coefficient of Determination) for the optimized SVM model was -0.0293, indicating that the model did not explain much of the variance in the data. A negative R^2 suggests that the model performed worse than the baseline, which simply predicts the mean of the sales.

The bar chart below compares the MSE between the Optimized SVM Model and the Baseline Model. The optimized SVM model showed a slight improvement over the baseline, but the MSE was still high, indicating room for model improvement.

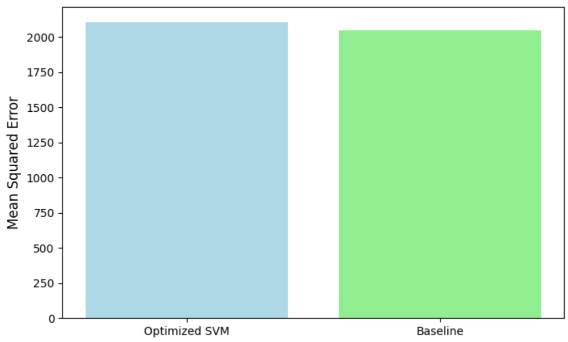

Figure 4. MSE Comparison: Optimized SVM vs Baseline Model

The scatter plot below compares predicted sales against actual sales. Ideally, the points would align along the **y = x** line, but the points are spread out, indicating that the model's predictions were not very accurate.

ISSN 2442-7888 (online) DOI 10.24167/sisforma.v12i2.14282

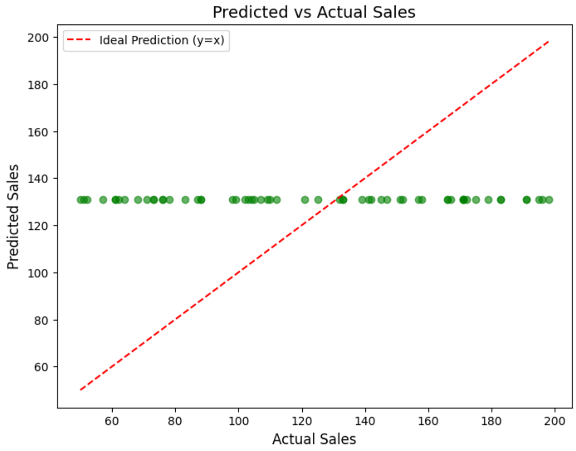

Figure 5 Predicted vs Actual Sales

The residual plot below visualizes the difference between actual and predicted sales. The residuals are scattered around the zero line without showing a clear pattern, suggesting that the model was not capturing significant patterns in the data.

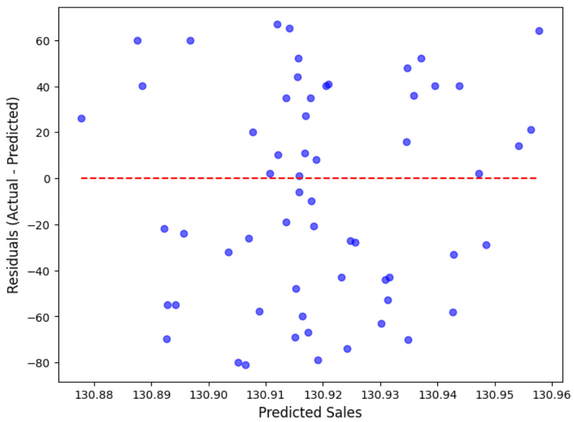

Figure 6 Residual Plot

B. DISCUSSION

The results of this study highlight several important insights into the use of Support Vector Machine (SVM) for sales prediction in Micro, Small, and Medium Enterprises (MSMEs). The MSE for the optimized SVM model was 2105.94, which indicates a high level of error in the model's predictions [17]. While the optimized SVM model performed better than the baseline model, the MSE remains relatively large. This suggests that the model struggles to capture the complexities of the sales prediction task [18]. The negative R^2 further indicates that the model failed to capture the variance in the data effectively. The negative R^2 value indicates that the model is underfitting the data, meaning that it is too simple to capture the underlying patterns in the data. The residual plot did not reveal any clear patterns, suggesting that the model is not overfitting but rather failing to learn the important relationships between features and target variables.

The features used in this model (price**,** stock, and promotion) are crucial but may not fully explain sales performance. Additional features such as seasonality, economic conditions, and market trends could significantly improve the model's ability to make accurate predictions [19]. While the polynomial kernel was chosen for its ability to capture non-linear relationships, further experiments with other kernels, such as Radial Basis Function (RBF) or linear, may yield better results. The selection of kernel directly influences the complexity of the model and its ability to generalize to unseen data [20]. The model can be further improved by: (a) Adding external variables such as seasonal effects, holiday promotions, or economic indicators could provide more predictive power, (b) Alternatives such as Random Forest or Gradient Boosting may be better suited for this type of prediction task, as they can handle complex relationships and interactions in the data more effectively, (c) Expanding the search space for C and gamma, or using techniques like RandomizedSearchCV, could improve model performance.

Despite the limitations, this model provides valuable insights for Micro, Small, and Medium Enterprises (MSMEs) in their digital transformation efforts [21]. By adopting predictive analytics, MSMEs can make more informed decisions related to inventory management, pricing, and promotional strategies [22]. Furthermore, this research provides a foundation for future improvements, including the incorporation of additional features, the exploration of alternative models, and the fine-tuning of model parameters.

In conclusion, while the model provides a basic framework for sales prediction, further refinement is necessary to achieve higher predictive accuracy. This research demonstrates the potential of machine learning in transforming sales forecasting for MSMEs, but there is still significant room for improvement and further investigation.

IV. CONCLUSION

This study successfully developed a predictive sales model for Micro, Small, and Medium Enterprises (MSMEs) using Support Vector Machine (SVM), optimized through GridSearchCV to identify the most effective combination of hyperparameters (C = 0.1, gamma = 0.1, and kernel = 'poly'). Despite improvements over the baseline model, as indicated by a reduced Mean Squared Error (MSE) of 2105.94, the model's performance remains suboptimal with an R^2 of -0.0293, signaling the challenges of capturing complex relationships in sales data for MSMEs.

This research contributes to the body of knowledge on applying machine learning, particularly SVM, in sales prediction for MSMEs—a topic that has been underexplored in existing literature. The findings suggest that while SVM holds potential for forecasting sales, additional features such as seasonal trends, economic indicators, and market fluctuations should be incorporated to enhance model accuracy.

The results also point to the limitations of the current SVM model, which struggles with underfitting, as evidenced by its negative R^2 and high residual errors. Future research should consider exploring alternative machine learning algorithms such as Random Forest or Gradient Boosting, which may handle complex feature interactions better. Further tuning of hyperparameters, as well as expanding the scope of data inputs, will likely improve prediction performance.

In conclusion, while this study provides a foundation for predictive sales modeling in MSMEs, there is considerable room for further refinement. The potential of machine learning to transform MSMEs' decision-making processes is evident, but future work should focus on overcoming the current limitations to achieve more reliable and actionable sales forecasts, thus aiding MSMEs in their digital transformation and long-term growth.

ACKNOWLEDGMENT

We would like to express our deep gratitude to Universitas Duta Bangsa Surakarta for the support and resources provided throughout the course of this research. Our sincere thanks also go to the community partners and respondents for generously dedicating their time and sharing valuable insights, which were instrumental in the success of this study.

Additionally, we extend our heartfelt appreciation to the Ministry of Higher Education, Science, and Technology (Kemendiktsaintek) for their crucial financial support, which was pivotal in enabling the successful completion of this research. Without their funding, this study would not have been possible.

REFERENCES

[1] F. O. Ugbebor, D. A. Adeteye, and J. Ugbebor, "Predictive analytics models for smes to forecast market trends, customer behavior, and potential business risks," *J. Knowl. Learn. Sci. Technol.*, vol. 3, no. 3, pp. 355–381, 2024, doi: 10.60087/jklst.v3.n3.p355-381.

[2] S. Evanita and Z. Fahmi, "Analysis of Challenges and Opportunities for Micro, Small, and Medium Enterprises (MSMEs) in the Digital Era in a Systematic Literature Review," 2023, doi: 10.32503/jmk.v8i3.4190.

[3] A. Mosbah, M. A. M. Ali, and N. Tahir, "Empowering Small and Medium Enterprises with Data Analytics for Enhanced Competitiveness," pp. 338–342, 2023, doi: 10.1109/iccsce58721.2023.10237151.

[4] S. Brandy, "Overcoming Challenges and Unlocking the Potential:

Empowering Small and Medium Enterprises (SMEs) with Data Analytics Solutions," 2023, doi: 10.58776/ijitcsa.v1i3.47.

[5] M. Gafarov, "Predictive Analytics for Sales Forecasting and Inventory Management," *Next Gener. J. Young Res.*, vol. 8, no. 1, p. 109, 2024, doi: 10.62802/7t6wq430.

[6] G. Ocran, S. O. Yusuf, P. Owusu, E. A. Boateng, S. O. Krampah, and A. H. Paul-Adeleye, "AI-driven business analytics for SMES: Unlocking value through predictive and prescriptive analytic," *Int. J. Sci. Res. Arch.*, vol. 13, no. 1, pp. 3009–3022, 2024, doi: 10.30574/ijsra.2024.13.1.2001.

[7] I. V Pustokhina and D. A. Pustokhin, "A Comparative Analysis of Traditional Forecasting Methods and Machine Learning Techniques for Sales Prediction in E-commerce," *Am. J. Bus. Oper. Res.*, vol. 10, no. 2, pp. 39–51, 2023, doi: 10.54216/ajbor.100205.

[8] H. Upadhyay, S. Shekhar, A. Vidyarthi, R. Prakash, and R. Gowri, "Sales Prediction in the Retail Industry Using Machine Learning: A Case Study of BigMart," in *2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM)*, 2023, pp. 1–6. doi: 10.1109/elexcom58812.2023.10370313.

[9] T. Gaertner, C. Lippert, and S. Konigorski, "Automated Demand Forecasting in small to medium-sized enterprises," pp. 1–14, 2024, [Online]. Available: http://arxiv.org/abs/2412.20420

[10] S. P. Lesmarna *et al.*, "Development of time-series-based mlops architecture for predicting sales quantity in micro, small, and medium enterprises (msmes)," *Transm. J. Ilm. Tek. Elektro*, vol. 26, no. 2, pp. 64–69, 2024, doi: 10.14710/transmisi.26.2.64-69.

[11] "Model for Predicting Prospective Big-Mart Sales Based on Grid Search Optimization (GSO)," 2022, doi: 10.1109/aide57180.2022.10059929.

[12] N. Kumar, "A Fine Tuned-based Framework to Predict Salesforce Data using Machine Learning in Business Analytics," *Eng. Technol. Appl. Sci. Res.*, vol. 14, no. 6, pp. 18563–18568, 2024, doi: 10.48084/etasr.8948.

[13] V. R. Sadasivam, G. Narendharakumar, M. Ragavan, and A. Prabhuram, "Food Sales Analysis and Prediction Using Machine Learning," pp. 1–5, 2024, doi: 10.1109/aimla59606.2024.10531342.

[14] I. G. Nalcaci, S. A. Satıcı, F. Alagöz, B. Onay, and C. Oksuz, "Enhancing Sales Prediction Accuracy Through Incorporation of Expert Knowledge as Constraints Into the Model," pp. 1–5, 2024, doi: 10.1109/aict61888.2024.10740302.

[15] A. M. Abdul-Yekeen, M. A. Kolawole, B. Iyanda, and H. A. Abdul-Yekeen, "Leveraging predictive analytics to optimize sme marketing strategies in the us," *J. Knowl. Learn. Sci. Technol.*, vol. 3, no. 3, pp. 73–102, 2024, doi: 10.60087/jklst.vol3.n3.p73-102.

[16] W. Wang, "Regression-Based Hyperparameter Learning for Support Vector Machines," *IEEE Trans. neural networks Learn. Syst.*, vol. PP, 2023, doi: 10.1109/tnnls.2023.3321685.

[17] H. Hua, X. Cui, F. Yuqiang, X. Ling, and W. Hao, "Product sale prediction method based on support vector machine model with parameter optimization." 2018.

[18] P. Ganguly and I. Mukherjee, "Enhancing Retail Sales Forecasting with Optimized Machine Learning Models," in *2024 4th International Conference on Sustainable Expert Systems (ICSES)*, 2024, pp. 884–889. doi: 10.48550/arxiv.2410.13773.

[19] H. Sharma *et al.*, "Enhancement of Sales Forecasting and Prediction with Machine Learning Methods," *Int. Res. J. Comput. Sci.*, vol. 11, no. 11, pp.

641–644, 2024, doi: 10.26562/irjcs.2024.v1111.02.

[20] T. Jin, "Optimizing Retail Sales Forecasting Through a PSO-Enhanced Ensemble Model Integrating LightGBM, XGBoost, and Deep Neural Networks," 2025, doi: 10.20944/preprints202501.1604.v1.

[21] K. Liu, "A Three-dimensional Model of Digital Transformation for Small and Medium-sized Enterprises," vol. 3, no. 2, pp. 112–118, 2023, doi: 10.54097/jid.v3i2.9396.

[22] S. G. Krishnan, "The Impact of Data Analytics Adoption on Job Roles and Skill Requirements in MSMEs: Development of Conceptual Model," *Int. J. Humanit. Educ. Soc. Sci.*, vol. 4, no. 3, 2024, doi: 10.55227/ijhess.v4i3.1313.