Asian Journal of Innovation and Policy (2025) 14.3:251-277
DOI: https://doi.org/10.7545/ajip.2025.14.3.251

Measuring the Impact of Informatization on SMEs Business Processes : Based on a Machine Learning Analysis Approach

Yoonseo Won*, Jongtae Lee**

Abstract This study aims to suggest a machine learning-based analysis model to predict the business process effectiveness and to analyze the informatization level survey data of small and medium-sized enterprises (SMEs). The study also focuses on identifying the most effective methodology and providing practical insights for establishing informatization strategies. This study predicts the business process effectiveness of small and medium-sized enterprises (SMEs) by utilizing the survey data of their informatization level. Representative machine learning classification models - Random Forest, XGBoost, and LightGBM - were applied, and SHAP (SHapley Additive exPlanations) was then used to analyze the key variables influencing the prediction performance with the highest-performance analysis model. According to the findings, the Random Forest and LightGBM models demonstrated the best performance in terms of AUC and accuracy for predicting business process effectiveness. A SHAP analysis suggested that the informatization capabilities of enterprises and the use of appropriate information systems could be key variables for business process effectiveness.

Keywords Business Informatization, Business Process Effectiveness, Machine Learning, SHAP(SHapley Additive exPlanations), Small and Middle Enterprise

Submitted, October 12, 2025; 1st Revised, November 30, 2025; Accepted, December 26, 2025
* Undergraduate Student, Department of Japanese Language & Literature, Seoul Women's University, Seoul, Republic of Korea; wonyyysss21@swu.ac.kr
** Corresponding, Associate Professor, Department of Business Administration, Seoul Women's University, Republic of Korea, light4u@swu.ac.kr

I. Introduction

1. Research backgrounds and purpose

Small and medium-sized enterprises (SMEs) play a crucial role in the national economy in terms of business scale and job creation. According to the report of the Korean Statistical Information Service (KOSIS) in 2022, the number of SMEs in Korea (based on business entities) amounted to approximately 8.04 million, representing 99.9% of all firms and enterprises, with about 18.96 million employees, accounting for 81% of the total workforce. However, many SMEs face difficulties surviving in stiff competition due to their limited scale and encounter constraints in achieving sustainable growth and maintaining employment (Han et al., 2013; Kim and Kim, 2015).

Lim (2002) suggests that it can also be defined as the degree of conversion of core business processes using digital technology, including advanced data analytics and information utilization. Informatization at the corporate level is generally understood as the extent of planning and evaluation for digital transformation (Viktor et al., 2020). Informatization has emerged as a critical factor for SMEs to secure competitive advantages and enhance their survival prospects. It refers to the activities through which firms use computers and information technologies (IT) to collect, process, store, and manage necessary information and apply it to decision-making, thereby improving the efficiency, effectiveness, and productivity of their business operations. This concept extends beyond the adoption of specific technologies and encompasses IT-based information activities, as well as the technological and organizational processes required to build the infrastructure that enables such activities (Jeong and Lee, 2024; Kim and Kim, 2015). However, according to the Korea Technology and Information Promotion Agency for SMEs (TIPA) and prior studies, many Korean SMEs face considerable challenges in establishing and utilizing informatization strategies. Even when a certain level of information systems is in place, limitations in investment, such as shortages of qualified professionals and constraints related to system maintenance and post management, impede the continuous operation and expansion of these systems (Park and Kwahk, 2020; Han et al., 2013; Sun, 2022). To mitigate these challenges, the Korea Technology and Information Promotion Agency for SMEs (TIPA) has implemented various support programs, including digital management systems and production automation projects, and has conducted annual assessments of firms' informatization levels. Nevertheless, due to budgetary limitations, the number of SMEs that can benefit from these programs remains only a small fraction of all firms, representing a significant constraint in terms of support coverage (Lee and Lee, 2016). These structural limitations, comprising both

internal operational and investment constraints, contribute fundamentally to widening informatization gaps among firms and ultimately hinder improvements in performance and competitiveness. In particular, because firms differ considerably in their levels of informatization and capabilities to utilize information systems, the effects of informatization are also likely to vary across firms. Consequently, the need to empirically identify how informatization influences actual business processes has become increasingly important. In this context, research that predicts and evaluates the impact of informatization levels on the effectiveness of business processes can provide essential evidence for establishing informatization strategies for SMEs and informing policy support directions.

This study focuses on predicting and evaluating business process effectiveness and seeks to resolve the uncertainty in informatization impacts arising from differences in informatization levels and utilization capabilities across firms. To this end, it analyzes the use of machine learning–based classification models that comprehensively consider multiple factors related to SME informatization. For this purpose, the "2020 Survey on the Informatization Level of SMEs" data, conducted by TIPA, was employed. This data includes responses from firms' practitioners evaluating their firms' informatization levels and business process effectiveness. To ensure analytical accuracy, this study dichotomized the evaluation of business process effectiveness into positive and negative categories and applied the data to prediction models based on representative machine learning classification methodologies of Random Forest, XGBoost, and LightGBM. The prediction performances of these models were compared based on accuracy indicators, and the model demonstrating the highest performance was selected. Subsequently, SHAP analysis was conducted on the selected model to identify significant informatization factors predicting business process effectiveness and to derive practical implications.

The structure of this paper, aligned with the research objectives, is as follows. Chapter 2 reviews prior studies relevant to this research, and Chapter 3 describes the research design and analytical procedures. Chapter 4 presents the empirical results, followed by Chapter 5, which discusses implications, limitations, and directions for future research.

II. Literature Reviews & Research Design

1. SME Informatization

In this era, knowledge and information are recognized as critical economic resources. In this regard, many firms have adopted informatization strategies to survive in intense competition and pursue sustainable growth, thereby seeking to enhance managerial efficiency and secure competitive advantage (Kim and Kim, 2015). SMEs have paid attention to informatization improving managerial efficiency and numerous studies have focused on SME informatization issues from various perspectives (Jeong and Lee, 2024). Kim et al. (2008) defined SME informatization as "The effectiveness of information technology in enhancing and utilizing SMEs' business processes" and Hahm and Kim (2023) emphasized that the degree of information system utilization can be related to a firm's business performance.

Previous studies have consistently reported that the higher level of informatization positively affects corporate performance including process innovation, productivity enhancement, prompt and efficient decision-making, and faster customer response (Sun, 2022; Hahm and Kim, 2023). In particular, the utilization of information systems has been considered as improving internal processes and facilitating effective communication within and between organizations (Won and Lee, 2021; Sun, 2022). van Zyl et al. (2022) emphasized that systematic accumulation and knowledge sharing using information systems can take place within SMEs and enable their prompt and accurate decision-making. This suggests that smoother information flows can enhance the overall effectiveness of business processes and highlights the necessity of managerial recognition and strategic support for informatization strategies including the adoption and implementation of information systems. These previous studies collectively indicate that informatization is closely associated with business processes, thereby emphasizing the need to predict business process effectiveness based on the informatization level of SMEs and to analyze the key contributing factors.

2. Business Processes and Process Effectiveness

A business process refers to a series of activities undertaken by firms to produce products or deliver services and consists of logically connected tasks to achieve business outcomes (Kang et al., 2008). Dumas et al. (2018) defined a business process as a structured collection of activities or tasks that produce a specific service or product for customers and explained it as the way organizations coordinate work, data, and information to create customer value.

Rosemann and vom Brocke (2010) argued that processes should not be regarded merely as isolated technical outputs but as strategic assets contributing to business success. Firms can increase profitability, maintain operational efficiency at lower costs, and leverage processes as core assets. With the growing diversity and rapid shifts in market considering customer demands, the ability to respond flexibly and to change swiftly has become crucial for achieving competitive advantage (Hong et al., 2008). To effectively manage uncertainties in the diverse market environment, firms must be able to react with agility and discern emerging opportunities. Agile business processes support this capability while also playing a key role in enabling firms to achieve cost economies (Roberts and Grover, 2012; Chen et al., 2014). Information technology capabilities can play a critical role in securing business process agility because higher utilization levels of information system utilization can contribute to both the flexibility and effectiveness of business processes simultaneously (Hahm and Kim, 2023; Sun, 2022). Collectively, these previous studies suggest that, in order to cope with market uncertainty and volatility, firms must strategically manage their business processes, strengthen process flexibility and effectiveness, and enhance their informatization capabilities to utilize information systems efficiently.

Through these activities, various elements interact to advance organizational goals to manage resources efficiently and effectively to sustain their existence (Durand and Vargas, 2003; Hahm and Kim, 2023). Efficiency refers to producing better output with the same resources or achieving the same output with less resources. Effectiveness, in contrast, reflects the degree to which the given means are appropriate for achieving objectives and the likelihood of success. Thus, assessing the extent to which processes enable firms or organizations to achieve their goals would be crucial (Burches and Burches, 2020).

Whereas previous studies have primarily focused on efficiency, from a business perspective, the ultimate objective of firms lies in generating long-term outcomes rather than merely reducing costs. Cho (2018) emphasized that effectiveness in achieving outcomes must be considered beyond efficient execution for corporate activities to translate into performance and that simply conserving resources to produce more outputs is insufficient unless such outputs are directly connected to firm objectives.

Regarding performance measurement of business processes, Van Looy and Shafagatova (2016) highlighted the importance of adopting multidimensional indicators such as time, cost, quality, and flexibility and explained that effectiveness is closely linked to organizational performance and strategic alignment. Del-Río-Ortega et al. (2018) introduced Process Performance Indicators (PPIs) enabling quantitative measurement of both efficiency and effectiveness. Their study underlines the central role of business process

effectiveness in evaluating the contribution of processes to organizational objectives.

In summary, business process effectiveness is a concept that must be considered in addition to efficiency, reflecting its contribution to achieving strategic goals. In this study, "decision-making speed" and "knowledge sharing" are identified as core sub-dimensions. These two sub-dimensions are supported by previous studies, which suggest that faster and more comprehensive decision-making systems enhance overall firm performance, and that knowledge sharing, both internal and external, contributes significantly to performance creation (Wang and Wang, 2012; Alzghoul et al., 2022). Therefore, these two factors serve as key criteria for evaluating how effectively business processes contribute to organizational objectives and can be regarded as core components of business process effectiveness.

3. Studies on Machine Learning–Based Prediction

In recent years, previous prediction studies adopting machine learning techniques have gained momentum in business administration and other social science fields. Compared with traditional methodologies used in the social sciences, machine learning has been demonstrated in previous studies to provide higher accuracy as well as robust classification and prediction models (Kim et al., 2023; Jeong and Lee, 2024).

Based on these academic advantages, numerous previous prediction studies in the field of business administration have applied machine learning yielding relatively higher predictive accuracy and performance (Yoon et al., 2022). For example, Moon et al. (2020) developed prediction models for crowdfunding success in Korea using machine learning algorithms such as Decision Tree, Gradient Boosting, and Random Forest. An and Lee (2021) employed ensemble techniques such as XGBoost and LightGBM to predict innovation outcomes and managerial performance resulting from corporate innovation. Yoon et al. (2022) proposed models for classifying and predicting firms in terms of management performance using Decision Tree, Support Vector Machine (SVM), and LightGBM. Jeong and Lee (2024) also introduced a comparison study employing Random Forest, XGBoost, and Feed-forward Neural Network (FNN) to predict business process efficiency based on SME informatization levels. In addition, Jang et al. (2023) proposed a smart consulting methodology for SMEs facing financial constraints in accessing management consulting, applying machine learning algorithms of Random Forest, XGBoost, and K-Nearest Neighbor (KNN). Table 1 summarizes the machine learning models and their performance outcomes reported in these major previous studies. Building on

these studies, the present research proposes prediction models for business indicators utilizing machine learning methodologies.

Table 1. Studies on Machine Learning–Based Prediction

Author	Machine Learning Model	Accuracy
Moon et al.(2020)	Decision Tree	85.17%
	SVM	86.02%
	NaiveBayes	87.77%
	RandomForest	88.34%
	AdaBoost	89.57%
	Gradient Boost	90.30%
An and Lee(2021)	XGBoost	93.02%
	LightGBM	93.94%
	CatBoost	90.97%
Yoon et al.(2022)	Logistic Regression	78.02%
	Decision Tree	77.84%
	SVM	76.58%
	RandomForest	79.58%
	LightGBM	77.12%
Jeong and Lee(2024)	XGBoost	95%
	RandomForest	95%
	Decision Tree	94%
	FNN	90%
	NaiveBayes	69%
Jung et al.(2023)	SGD	76%
	KNN	79%
	RandomForest	85%
	XGBoost	86%

lll. Research Model and Variable Definition

1. Research Model

The explanatory variables were constructed with reference to the variables selected through logistic regression in the study of Jeong and Lee (2024). These variables include firm type, general firm information (industry, sales, type of

R&D activity, etc.), commitment and plans for informatization, informatization implementation environment, degree of information system development, utilization level, and informatization effectiveness. These explanatory variables are applied to machine learning–based classification models to predict whether the informatization level of a firm exerts a positive or negative impact on business process effectiveness.

The analytical procedure of this study consists of six steps: (1) collection of SME informatization survey data, (2) data preprocessing, (3) variable selection, (4) classification modeling, (5) model performance evaluation, and (6) SHAP analysis.

First, in the data collection stage, this study adopted the data from the '2020 Survey on the Informatization Level of SMEs' provided by the Korea Technology and Information Promotion Agency for SMEs under the Ministry of SMEs and Startups. Second, during data preprocessing, missing values and outliers were removed. Following previous studies indicating that a missing rate of 20–30% is generally acceptable in the information systems field, data with missing values below 30% were used to ensure analytical stability (Peng et al., 2022). To facilitate accurate classification, the dependent variable was discretized into 'positive contribution to business process effectiveness' and 'negative contribution to business process effectiveness.'

Third, in the variable selection stage, this study adopted the key explanatory variables identified in the study of Jeong and Lee (2024) to avoid prolonged training time and potential errors from using the full dataset. Additional variables suggested in previous studies were considered to align with the objectives of the present study.

Fourth, in the classification modeling stage, three representative machine learning–based classification models—Random Forest, XGBoost, and LightGBM—were applied and their performances were evaluated carefully.

Finally, SHAP analysis was conducted to assess the contribution of the key variables influencing prediction results and practical implications for SME informatization strategy formulation were derived based on these findings.

2. Description of Data

This study utilized the '2020 Survey on the Informatization Level of SMEs' provided by the Korea Technology and Information Promotion Agency for SMEs (TIPA) under the Ministry of SMEs and Startups to predict business process effectiveness based on informatization levels. The data employed in this study were collected to assess and evaluate the informatization status and digital maturity level across different informatization domains of domestic SMEs and have been gathered annually since 2000.

The 2020 survey was conducted over approximately six weeks including the data collection and compilation period. The survey result covers various industrial sectors including manufacturing (food and beverages, textiles/ apparel, petroleum/chemicals, machinery/metals, electronics, and other manufacturing businesses), construction, wholesale and retailing, and transportation. The survey replies consist of 4,000 SMEs, 300 large enterprises, and 300 supported firms, with the 4,000 SMEs stratified according to seven revenue categories and twelve industry types.

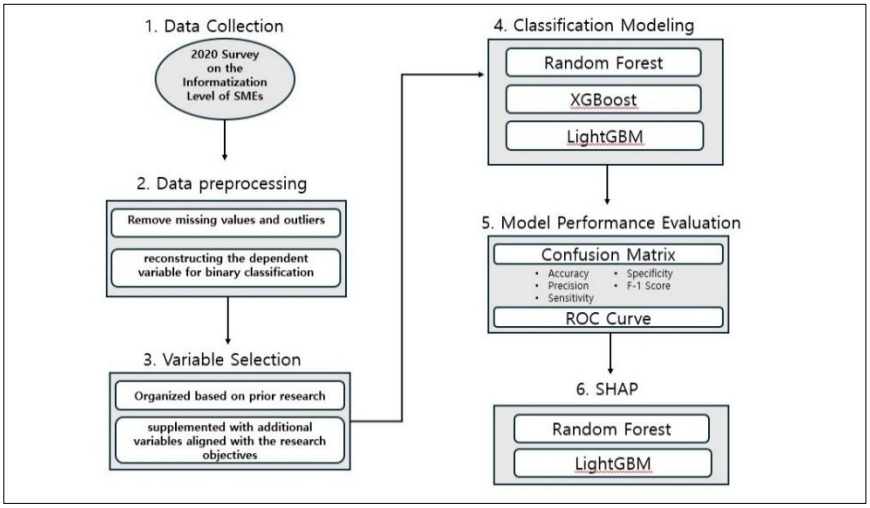

Figure 1. Research Model

The questionnaires in the survey were designed to comprehensively assess the informatization and digital transformation levels of SMEs, encompassing six main domains ranging from general firm information to informatization effectiveness, as well as the adoption status of smart factories and ICT-related advanced technologies. The detailed items for each domain are demonstrated in Table 2.

Table 2. Questionnaire Structure

Domain	Detailed items
A: General Firm Information	Revenue size, industry, organizational form, firm type (as of 2020), whether located in an industrial complex, year of establishment, number of employees, number of female employees, current number of full-time employees, company's industry, type of R&D activities, capital, sales, operating profit, whether the firm engages in exports, main product types, primary sales channels by revenue, and the most significant customer among the sales channels
B: Commitment and Plans for Informatization	interest in informatization among organizational members (top executives, managers, and employees), the extent to which informatization implementation plans have been established, and the analysis of the feasibility of informatization investments
C: Informatization Implementation Environment	Understanding of internal informatization tasks and information, system environment, Informatization investment costs, Informatization training, Informatization personnel and outsourcing, Speed of informatization-driven business process innovation, Information security systems, Information system management (post-management and maintenance)
D: Information System Development and Utilization Status	Currently used services among mobile office, cloud, and SNS Plans to expand the scope of service usage, Information system implementation status (e.g., sales, marketing, customer management processes, sales planning, electronic procurement systems), Level of inter-firm process utilization
E: Informatization Effectiveness	Proportion of business operations, Contribution of information systems to operational efficiency, Contribution of information systems from the BSC (Balanced Scorecard) perspective
F: Smart Factory and ICT Advanced Technology Adoption	Adoption status of smart factories and ICT advanced technologies

3. Definition of the Variables

In this study, the dependent variable, business process effectiveness, was constructed by calculating the average of the responses to the questionnaire items on "knowledge sharing in work" and "speed of decision-making." Both items were measured on a 5-point Likert scale: "Very High," "High," "Moderate," "Low," and "Very Low." Combining similar variables into a single composite variable using the mean or weighted mean is a commonly employed method to enhance analytical efficiency and interpretability (Song et al., 2013).

For the definition of the dependent variable, responses of "Very High" and "High" were coded as 1, while "Low" and "Very Low" were coded as 0. Responses marked as "Moderate" were excluded, as they were considered

neutral and could introduce ambiguity when classifying business process effectiveness into positive or negative categories. In addition, given that the responses are based on subjective judgment, there may be inconsistencies in distinguishing between "Very Low" and "Low" or "High" and "Very High" among different respondents under identical circumstances. Therefore, similar responses were grouped together as described above. This recoding process transformed the dependent variable from a multi-class problem into a binary classification problem. After this preprocessing, the variable could be relatively less complex and expected to yield higher performance metrics. Accordingly, a binary classification model was developed to categorize business process effectiveness as positive or negative. Although 4,000 SMEs participated in the informatization survey, after data preprocessing to remove missing and outlier values, a total of 1,309 valid responses were used for analysis.

For the selection of explanatory variables to predict business process effectiveness based on the informatization level of SMEs, this study referred to the previous studies by Jeong and Lee (2024). In their study, logistic regression analysis which is commonly used for binary classification problems and can yield statistically significant results was applied and variables with p-values below 0.05 were included as explanatory variables. Based on the questionnaire structure and the actual response data, this study referred to variables suggested in relevant previous studies and further added or certain restructured items to align with the objectives of the present study. Detailed information on the composition of explanatory variables is explained in Table 3.

Table 3. Composition of explanatory variables

Category	Question Code	Description	Reference
General Information	SQ5	Firm Type (as of 2020)	Jeong and Lee (2024)
	SQ6~SQ6T	Certification	
	QA01~QA01Z02	Industry	
	QA02~QA02T	Type of R&D Activities	
	QA0311~QA0312	Capital	
	QA0321~QA0322	Sales	
	QA0331~QA0332	Operating Profit	
	QA031	Export Performance	
Informatization Implementation Capability	QB01~QB01Z05	Overall interest in informatization of top executives, managers, and employees (interest in informatization, willingness to support informatization, promotion of informatization investment plans and strategy formulation, etc.)	
	QB02	Extent of informatization implementation plan formulation (e.g., ERP, office automation, logistics systems, etc.)	
Informatization Operational Capability	QC0111~QC0122T	New investments and maintenance	
	QC02~QC021Z03	Informatization training and target participants	
	QC03~QC034	Status of personnel in charge of informatization (including on-site outsourced staff) – dedicated staff, concurrent staff, outsourced management	Han et al. (2013)
	QC04~QC041Z01	Outsourcing ratio of informatization tasks	Jeong and Lee (2024)
	QC0411~QC0411T	Reasons for adopting (or planning to adopt/expand) informatization outsourcing	
	QC05	Business process innovation for informatization (changes in work methods and revision of related regulations and policies)	
	QC07101~ QC07306	Awareness and system level of information security (management of storage media, prevention of personal and internal information breaches, response to viruses and malware)	
	QC08~QC08Z02	Post-management and maintenance level of information systems (hardware, software, network)	

	QC081~QC081T	Causes of difficulties in post-management and maintenance	
Level of Informatization Utilization	QD01~QD02T	Use of mobile office (smart work) utilizing mobile devices and wireless internet, and plans for expansion	
	QD04101~QD04121	Implementation status (sales planning process / production and inbound planning process / financial processes such as budgeting, settlement, and cost management accounting / ERP, etc.)	
	QD04201~QD04221	Level of internal business process utilization (sales, marketing, customer management processes / e-payment / production process management / quality management processes / internal knowledge sharing systems (KMS, EIP, EKP, etc.) / SCM (Supply Chain Management systems), etc.)	
	QD04301~QD04310	Level of inter-firm utilization (e-procurement system / production and inbound planning processes / shipping planning processes / quality management processes, etc.)	
Information System Effectiveness	QE011~QE011Z04	Proportion by item (number of employees in sales, marketing, and customer management / proportion of tasks supported by information systems within total business operations)	
	QE021~QE021Z01	Assessment of the contribution of information systems to the efficiency improvement of each process (reduction of effort in executing business processes, improvement in accuracy of business processes)	Hahm and Kim (2023)
	QE023~QE023Z04	Assessment of the contribution of information systems to the improvement of process efficiency and effectiveness by business area (sales and purchasing management, production, development, etc.)	
	QE03~QE03Z05	Assessment of the overall contribution of information systems to corporate performance, including financial outcomes (sales, net profit, etc.), internal processes, customer satisfaction, and key performance indicators	Sun (2022)
Smart Factory and ICT Advanced Technologies	QF01~QF11Z01	Impact of smart factories, budget allocation for smart factory implementation, adoption of ICT advanced technologies such as cloud, big data, and AI, budget allocation for ICT technology adoption, and difficulties encountered during utilization	Jung et al.(2023), Fatoba et al.(2024)

IV. Research Methodology Design and Empirical Results

1. Classification Models

In this study, Random Forest, XGBoost, and LightGBM were applied and compared to suggest better classification models, and all analyses were conducted using R (version 4.5.0). The characteristics of each methodology are as follows.

Random Forest is a bagging-based approach proposed by Breiman (2001). It generates multiple sub-datasets by bootstrapping the training dataset and trains a decision tree on each sub-dataset. Final class predictions are determined through majority voting (Lee, 2024; Eom et al., 2020). By averaging predictions from multiple models created via bootstrapping, Random Forest effectively reduces bias and variance that may arise from overfitting. However, when all variables are used, correlations between individual models may increase, so the model is constructed using sampling without replacement (Cho and Kim, 2021). By randomly selecting predictors and observations during tree formation, independent decision trees are generated, reducing prediction errors and providing reliable predictive performance (Yoon et al., 2022).

XGBoost (eXtreme Gradient Boosting) is a kind of Gradient Boosting Machine (GBM)-based algorithm that combines relatively weak classifiers into a stronger predictive model, reducing variances, and enhancing overall prediction accuracy (Chen and Guestrin, 2016). During training, it minimizes learning loss while controlling overly complex models to prevent overfitting, using the Gini index as the impurity measure for classification. The model employs a greedy algorithm for splitting and computes optimal weights through distributed processing (An and Lee, 2021). XGBoost is highly scalable and can handle missing data efficiently, demonstrating optimal performance even with sparse datasets (Lee, 2024). Using Classification And Regression Trees (CART) in the ensemble allows training with both categorical and continuous variables and each leaf contributes to the final prediction enabling comparison of individual tree results. Additionally, unnecessary branches are pruned when improvements in the loss function fail to meet a threshold, preventing overfitting (Chen and Guestrin, 2016).

LightGBM (Light Gradient Boosting Machine) is a decision tree-based ensemble model developed by Microsoft to overcome the limitations of XGBoost constructing highly predictive models using a leaf-wise growth strategy (Ke et al., 2017). To accelerate data processing and training speed, LightGBM employs Gradient-based One-Side Sampling (GOSS), focusing on data points with larger gradients rather than using all data for split-point searches. Furthermore, the Exclusive Feature Bundling (EFB) algorithm reduces the

number of input variables by combining correlated features into a single feature. These approaches significantly improve training speed and reduce memory usage (An and Lee, 2021; Ke et al., 2017).

Tree-based ensemble models, such as Random Forest, XGBoost, and LightGBM, are generally robust to multicollinearity because they partition data using variable splitting criteria instead of parameter estimates. This characteristic of ensemble learning models may effectively mitigate instability, offering more reliable predictive performance than traditional statistical approaches (Kim and Lee, 2020).

2. Performance Evaluation

Table 4. Confusion Matrix

Category		Predicted Class	
		True(1)	False(0)
Actual Class	True(1)	True Positive	False Negative
	False(0)	False Positive	True Negative

To evaluate model performance, this study utilized the Receiver Operating Characteristic (ROC) curve, accuracy, precision, specificity, sensitivity, and F1-score.

The ROC curve is a representative metric for evaluating the predictive power of machine learning classification methods. It can illustrate the relationship between sensitivity and specificity at various threshold values and allow the assessment of classifier performance through changes in the threshold values. The area under the ROC curve (AUC) is commonly used, with values above 0.9 indicating excellent performance.

Table 5. Classification Model Evaluation Metrics

Evaluation Metrics	Formula	Evaluation Metrics	Formula
Accuracy	$\frac{TP+TN}{TP+FN+FP+TN}$	Sensitivity	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$	F-1 Score	$2\times\frac{Precision\ \times Sensitivity}{Precision+Sensitivity}$
Specificity	$\frac{TN}{FP+TN}$		

Complementary evaluation metrics for classification models are derived from the confusion matrix including accuracy, precision, specificity, sensitivity, and F1-score (An and Lee, 2021). Accuracy represents the proportion of correctly classified instances, while precision measures the proportion of instances predicted as True (1) that are actually True (1). Specificity is the proportion of instances that are actually False (0) and correctly classified as False (0). Sensitivity, also known as recall, is the proportion of instances that are actually True (1) and correctly classified as True (1). Finally, F1-score, a harmonic mean of precision and sensitivity, is particularly suitable for imbalanced datasets.

3. Model Analysis Results

In this study, 80% of the entire dataset was randomly selected as the training data, and the remaining 20% was used as the evaluation data to assess the predictive performance of the classification models. Since the main focus is on whether the dependent variable - business process effectiveness - has a positive impact, cases with a value of '1' in the dependent variable were designated as the positive class for the analysis.

Given the characteristics of machine learning, hyperparameter tuning is crucial for achieving optimal performance. Therefore, a Grid Search method was employed to explore various combinations of hyperparameters to enhance the predictive performance of the Random Forest, XGBoost, and LightGBM models (Lee, 2024).

For Random Forest, the grid search explored "mtry" values of 4, 5, and 6, and "ntree" values of 400, 500, and 600. For XGBoost, learning_rate was set to 0.01, 0.05, and 0.1; max_depth to 3 and 5; subsample and colsample_bytree to 0.8 and 1.0; min_child_weight to 1 and 5; and gamma to 0 and 0.1. For LightGBM, learning_rate was set to 0.01, 0.05, and 0.1; num_leaves to 15, 20, and 31; and both feature_fraction and bagging_fraction to 0.6, 0.7, and 0.8.

To ensure model stability and generalizability, 5-Fold cross-validation was conducted for all models.

Table 6. Hyperparameter

Model	Hyperparameter
RF	mtry = 4,5,6
	ntree = 400, 500, 600
	method = "cv“
	number = 5
	metric = “ROC”
	importance = TRUE
XGBoost	booster = "gbtree"
	objective = "binary:logistic“
	eval_metric = "auc"
	learning_rate = 0.01, 0.05, 0.1
	max_depth = 3~5
	subsample = 0.8~1.0
	colsample_bytree = 0.8~1.0
	min_child_weight = 1~5
	gamma = 0~0.1
	nrounds = 517
LightGBM	objective = "binary"
	metric = "auc"
	learning_rate = 0.01, 0.05, 0.1
	num_leaves = 15, 20, 31
	feature_fraction = 0.6, 0.7, 0.8
	bagging_fraction = 0.6, 0.7, 0.8
	bagging_freq = 5
	verbosity = -1
	seed = 123

Note: For other parameters not explicitly mentioned, default values were used.

The analysis results for each model are as following. First, comparing the accuracy of the models, LightGBM achieved the highest value at 0.87, followed by Random Forest and XGBoost, with all three models showing accuracy above 0.85 indicating strong performance in terms of accuracy. Considering that the data used are survey responses from the social sciences, these results suggest that the models are sufficiently applicable for classification and possess practical utility.

Table 7. Prediction Performance Evaluation

Model	Acc.	Pre.	Sen.	Spe.	F-1
RF	0.864	0.867	0.853	0.874	0.863
XGBoost	0.854	0.819	0.917	0.789	0.865
LightGBM	0.874	0.868	0.887	0.859	0.877

Regarding precision, LightGBM and Random Forest were measured at approximately 0.87 while XGBoost showed a slightly lower value of 0.82 - still maintaining a reasonable level of precision. This difference can be attributed to the fact that LightGBM applies efficient algorithms such as GOSS and EFB resulting in stable performance across accuracy and precision, whereas XGBoost is highly sensitive to parameter settings and, due to its boosting structure, tends to emphasize sensitivity over precision. Given that such patterns are commonly observed in case-based research, these differences can be interpreted as natural outcomes reflecting model characteristics.

Next, examining sensitivity (recall), XGBoost recorded the highest value at 0.92, followed by LightGBM and Random Forest. Specificity and F1-Score were generally above 0.85 across the models. Considering all performance metrics together, Random Forest and LightGBM demonstrate relatively strong performance, with minimal differences observed between the two models.

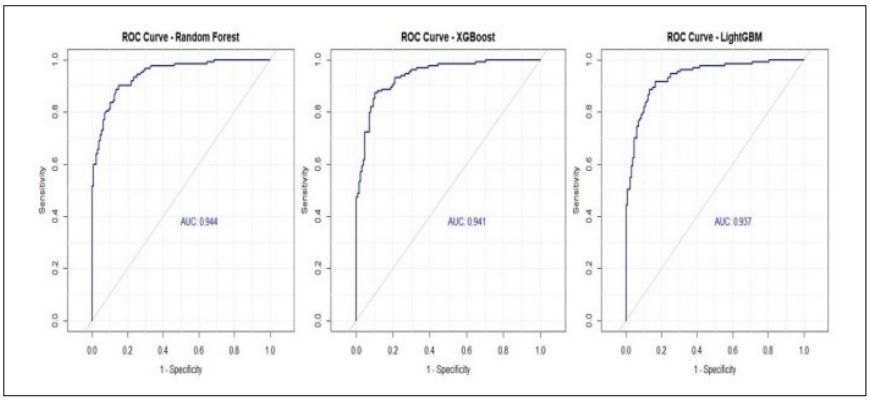

Figure 2. ROC Curves for Each Model

Next, examining the ROC curves of each classification model in Figure 2, all three models—RF, XGBoost, and LightGBM—demonstrate excellent classification performance with AUC values above 0.9 and close to 1. Numerically, RF shows the highest AUC at 0.944 followed by XGBoost at 0.941 and LightGBM at 0.937; the differences between the latter two models are less than 0.01 indicating no meaningful performance difference among the three models. While the ROC curve analysis suggests that RF and XGBoost perform relatively well, it is generally recommended to also consider evaluation metrics such as accuracy and precision for a more precise assessment of model performance. Considering both the ROC curves and evaluation metrics like accuracy and precision, RF—with an AUC of 0.944 and overall strong metric results—can be regarded as exhibiting the best performance. However, in terms of accuracy and precision, LightGBM records the highest values across almost all metrics while its AUC shows no significant difference from others. Based on this, RF and LightGBM are selected as the top-performing classification models.

Overall, the analysis shows that the three models—Random Forest, XGBoost, and LightGBM—all representative ensemble techniques exhibit generally superior performance compared to other methods, a finding supported by several previous studies and confirmed in this study (Moon et al., 2020). Interestingly, unlike many studies reporting higher predictive performance for GBM based models such as XGBoost and LightGBM, RF showed the best performance in this study.

4. SHAP Analysis

Machine learning based classification models such as RF and LightGBM exhibit excellent predictive performance; however, they have limitations and vulnerabilities in precisely explaining the relationships between the dependent and explanatory variables. To address this issue, related previous studies have been conducted, and methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanation) have been developed (Lundberg and Lee, 2017). In this study, SHAP analysis, which is effective for selecting key explanatory variables, was performed to identify the main factors influencing predictions and to provide practical, actionable insights for business applications. SHAP analysis explains the importance of each variable using the Shapley value, which quantifies the contribution of a specific variable to the dependent variable (Thiruvengadam et al., 2022). The Shapley value, a core component of SHAP analysis, is calculated based on the average difference in outcomes with and without the target explanatory variable across all possible combinations of variables capturing its influence on the prediction (Oh et al., 2020). In this study, SHAP analysis was conducted for RF, which exhibited the highest AUC, and LightGBM, which demonstrated the highest accuracy. Figure 3 presents the feature importance plots showing the impact of each explanatory variable on the dependent variable of business process effectiveness. The plot displays explanatory variables in descending order of contribution, with longer bars indicating variables that have a greater impact on the dependent variable.

First, examining the SHAP analysis results of RF, the variables QE023z01, QE03, QE03z01, QE023z02, and QE03z02 are shown to have the highest importance. The QE023 series of items represents “the impact of information systems on process performance across business areas,” while the QE03 series represents “the overall contribution of information systems to firm performance.” The key variables with substantial influence on the dependent variable are all related to information systems, indicating that information systems significantly contribute to business process effectiveness. In addition to these five variables that explain the impact of information systems on firm performance and processes, variables in the QD04 series, which measure the level of information system implementation and utilization, also show high importance. This suggests that not only implementing information systems but also effectively utilizing them meaningfully affects business process effectiveness.

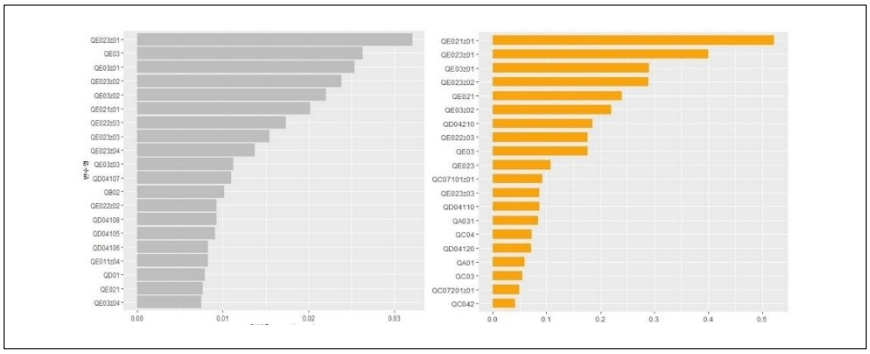

Figure 3. Comparison of SHAP Analysis Results between Random Forest (left) and LightGBM (right)

Next, examining the results of LightGBM, the variable importance is ranked QE021z01, QE023z01, QE03z01, QE023z02, and QE021. Overall, the results are similar to those of RF, but notably, the QE021 series shows higher importance. This variable represents "the degree to which information systems contribute to process efficiency improvement," indicating how accurately tasks can be performed and how effectively the time and effort required for task completion can be reduced to achieve the same results with fewer resources. Through efficient business processes, firms can make rapid and accurate decisions, facilitate systematic information sharing, and positively influence business process effectiveness. In addition to the five variables discussed earlier, variables from the QC03 and QC04 series in the informatization implementation environment, which describe the level of dedicated personnel and outsourcing, are also highlighted. These reflect the informatization capabilities and the utilization of external resources, all of which play a key role in the implementation and operation of information systems. These variables enable efficient execution of business processes, thereby making a significant contribution to overall business process effectiveness.

V. Conclusion

1. Research Summary

This study utilized the "2020 Survey on the Informatization Level of SMEs" to examine the impact of informatization on business process effectiveness and to develop predictive models using representative ensemble machine learning methodologies - Random Forest (RF), XGBoost, and LightGBM. SHAP

analysis was conducted on RF, which exhibited the highest AUC, and LightGBM, which showed the highest accuracy and precision to identify and analyze the key variables influencing the predictions. The summary of the study findings is as follows:

First, the literature review confirmed that informatization is regarded as enhancing managerial efficiency and positively affecting firm performance including process innovation and productivity improvement. Furthermore, when an informatization strategy is implemented, efficient and effective business process management can be achieved not only through the initial implementation of information systems but also via systematic operations such as raising awareness of informatization and providing strategic support.

Second, comparing ensemble machine learning methodologies, predictive models for business process effectiveness were analyzed and selected. The AUC values were 0.944 for RF, 0.941 for XGBoost, and 0.937 for LightGBM with RF demonstrating the highest performance in terms of AUC due to its robustness against noise. Meanwhile, LightGBM achieved the highest values across all metrics of accuracy and precision, all exceeding 0.85. This can be attributed to the characteristics of the survey data used, which may contain considerable noise due to subjective judgments and non-responses; bagging-rule based RF model is less sensitive to noise and thus delivers more robust performance (Bentéjac et al., 2021)

Third, SHAP analysis was applied to identify the key explanatory variables contributing to the dependent variable. The analysis of RF and LightGBM revealed that variables related to information systems had the greatest impact consistent with previous research indicating that the utilization of information systems significantly affects firm performance and business processes (Sun, 2022).

2. Academic and Practical Implications

This study proposes a predictive model for business process effectiveness using ensemble machine learning methodologies and evaluates its performance. The academic and practical implications derived from the study are as follows. The academic implications are, first, that this study reexamined the concept of business process effectiveness, which evaluates how effectively firms achieve the ultimate goal of enhancing firm performance using business processes presented in previous studies and confirmed that both efficiency and effectiveness can be quantitatively assessed (van Looy and Shafagatova, 2016; Ortega et al., 2018; Burches and Burches, 2020). Furthermore, based on several previous studies demonstrating the importance of business processes in directly influencing firm performance, this study extends its academic focus from an

efficiency-centered view to propose a model that predicts business process effectiveness with high accuracy highlighting its significance.

Second, diverse previous studies on informatization and business processes employed traditional methodologies, while machine learning studies faced the limitation of not being able to accurately explain causal relationships between explanatory and dependent variables. In this study, SHAP analysis is applied to select the key variables influencing the prediction of the dependent variable, and the results show that the utilization of information systems and firms' informatization capabilities significantly contribute to business process effectiveness to overcome the limitation. This aligns with previous studies demonstrating that information systems facilitate rapid decision-making and improve business processes through statistical analysis, thereby positively impacting firms' performance (Park and Kwahk, 2020; Won and Lee, 2021; Sun, 2022). Accordingly, this study empirically verifies that SHAP analysis can be used to address the limitations of previous machine-learning based studies in explaining relationships between variables.

Third, although machine learning techniques have been increasingly applied in management research, very few studies have directly predicted business process effectiveness, which was the objective of this study. This study compared the performances of the representative machine learning methodologies such as RF, XGBoost, and LightGBM to predict business process effectiveness with high accuracy, and by designing and evaluating machine learning-based classification models, it systematically confirms the validity of these predictive methodologies.

The practical implications are, first, that firms can use the proposed model to predict business process effectiveness based on their informatization data, aiding in strategic decision-making and identifying areas for improvement. The model's high predictive performance also supports the suggestion of the prediction model proposed by Jeong and Lee (2024) and allows firms to assess whether their business processes align with organizational goals. The machine learning framework proposed in this study enables public institutions to identify firms with high potential for performance improvement when allocating informatization support programs. At the same time, firms themselves can understand the key factors influencing effectiveness and improve deficient areas through continuous management, thereby highlighting the academic value of this approach.

Second, through SHAP analysis, key variables with high importance and contribution to business process effectiveness were identified. The results indicate that the operational and utilization levels of information systems significantly influence process effectiveness. Firms can use these findings to prioritize critical factors when implementing informatization, optimize processes, and enhance related capabilities. For example, firms could regularly

provide information system training and establish post-implementation management systems to continuously improve system utilization.

Third, given the model's high predictive accuracy and flexibility, SMEs can leverage the model to test adjustments to explanatory variables when business process effectiveness is predicted to be negative, thereby exploring conditions that may yield positive outcomes and establishing more effective informatization strategies.

3. Research Limitations and Future Study Suggestions

Despite the academic and practical implications, this study still has several limitations. First, since secondary survey responses were used, unexpected and inevitable subjective judgment may have affected the answers, and errors or omissions could have led to inaccurate data. Second, although SHAP analysis identified key variables for predicting the dependent variable, it remains challenging to precisely explain causal relationships among variables. Third, the study used cross-sectional data from 2020, limiting the ability to analyze trends or changes over time.

For future research, it would also be meaningful to consider the following topics. First, although this study only conducted SHAP analysis, it would be meaningful for future studies to conduct more in-depth analyses using a variety of methods to explain the causal relationships between explanatory variables and the dependent variable accurately. Second, this study confirms that the utilization of information systems significantly contributes to business process effectiveness. We propose that future studies can explore the key factors in the implementation and operation of information systems and analyze their structural relationships. Third, while this study was based on cross-sectional data from a single point in time, future studies can collect multi-year data and analyze the data to understand temporal changes or trends providing more in-depth findings.

References

Alzghoul, A., Khaddam, A.A., Abousweilem, F., Irtaimeh, H.J., and Alshaar, Q., (2022), How business intelligence capability impacts decision-making speed, comprehensiveness, and firm performance, Information Development, 40, 2, 220–233.

An, K., and Lee, Y., (2021), Corporate Innovation and Business Performance Prediction Using Ensemble Learning. The Journal of Information Systems, 30(4), 247-275. **

Bentéjac, C., Csörgő, A., and Martínez-Muñoz, G., (2021), A comparative analysis of gradient boosting algorithms, Artificial Intelligence Review, 54, 3, 1937-1967.

Breiman, L., (2001), Random forests, Machine learning, 45, 1, 5-32.

Burches, E., and Burches, M., (n.d.), Efficacy, Effectiveness and Efficiency in the Health Care: the Need for an Agreement to Clarify Its Meaning, International Archives of Public Health and Community Medicine, 4, 1, 1-3.

Chen, T., and Guestrin, C., (2016), XGboost: A Scalable Tree Boosting System, In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

Chen, Y., Wang, Y., Nevo, S., Jin, J., Wang, L., and Chow, W.S., (2014), IT Capability and Organizational Performance: The Roles of Business Process Agility and Environmental Factors, European Journal of Information Systems, 23, 3, 326-342.

Cho, K. and Kim. Y., (2021), Comparison of bankruptcy prediction models using statistical learning at multiple times. Journal of the Korean Data And Information Science Society, 32(3), 487-499. 10.7465/jkdi.2021.32.3.487 **

Cho, Y., (2018), Assessing the R&D Effectiveness and Business Performance: A Review of Their Mechanisms and Metrics, STI Policy Review, 9, 1, 1-29.

del-Río-Ortega, A., Resinas, M., and Ruiz-Cortés, A., (2018), Business Process Performance Measurement, In: Sakr, S., Zomaya, A. (eds.), Encyclopedia of Big Data Technologies, Springer, Berlin, 416-422.

Durand, R., and Vargas, V., (2003), Ownership, organization, and private firms' efficient use of resources, Strategic Management Journal, 24, 7, 667-675.

Eom, H., Kim, J., and Choi, S., (2020), Machine learning-based corporate default risk prediction model verification and policy recommendation: Focusing on improvement through stacking ensemble model. Journal of Intelligence and Information Systems, 26(2), 105-129. **

Fatoba, T.M., Jaiyeoba, G., Oladosu, O.T., and Oyewole, M.O., (2024), The Effect of Smart Factory on the Continuous Improvement of the Production Process: A Review, International Journal of Engineering and Modern Technology, 10, 1, 83-107.

Hahm, Y. and Kim, A., (2023), A Study on Business Performances According to the Informatization Environment of the Small and Medium Size Enterprises. Journal of Product Research, 41(1), 7-14. **

Han, H., Kim, K., and Yang, H., (2013), SME Informatization Attributes Based Analysis for their Criticalness, Status and Policy Implications. Journal of Information Technology Applications and Management, 20(4), 97–110. https://doi.org/10.21219/JITAM.2013.20.4.097. ** 21219

Hong, M., Moon, M., and Yeom, K., (2008), An Automatic Business Process Model Generation Tool Using Business Process Family Models. Journal of KIISE : Software and Applications, 35(8), 479-492. **

Jeong, J. and Lee, J., (2024), Machine-learning-based Prediction Model of Business Process Efficiency based on Informatization Level of SMEs. Journal of The Korean Operations Research and Management Science Society, 49(2), 49-69. **

Jung, S., Kim, D., and Shin, N., (2023), Success Factors of the Adoption of Smart Factory Transformation: An Examinitation of Korean Manufacturing SMEs, IEEE Access, 11, 2239-2249.

Kang, S., Lee, D., Lee, J., Yim, H., and Ahn, Y., (2008), Measuring Value Achievement in Business Processes, The KIPS Transactions : Part D, 15(3), 337-346.**

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W, and Liu, T.Y., (2017), Lightgbm: A Highly Efficient Gradient Boosting Decision Tree, Advances in Neural Information Processing Systems, 30, 3146-3154.

Kim, I., and Lee, K.,(2020), Tree based ensemble model for developing and evaluating automated valuation models: The case of Seoul residential apartment. Journal of the Korean Data & Information Science Society, 31(2), 375–389. **

Kim, J., and Kim, H., (2015), Design of Information Systems Audit Model for the Small and Medium Enterprise's Informatization Level Evaluation. Journal of Information Technology Services, 14(4), 105-120. **

Kim, Y., Cho, H., Lee, C., and Cho, J., (2023), A Study on the Machine Learning Model for the Financial Performance Prediction of Startups. Asia-pacific Journal of Convergent Research Interchange, 9(7), 67-77. **

La Rosa, M., Mendling, J., Reijers, H.A., and Dumas, M., (2018), Fundamentals of Business Process Management, Springer, Berlin.

Lee, H., and Lee, O.,(2016), Study on the Informatization Policy Evaluations and Directions for Small and Medium Enterprises (SMEs). Journal of the Korea Academia-Industrial cooperation Society, 17(10), 655-665. **

Lee, J., (2024), Exploration of Digital Divide Factors Affecting Life Satisfaction in Older People: Using Tree-Based Ensemble Machine Learning and SHAP. Journal of Digital Contents Society, 25(7), 1847-1860. **

Lim, S.K. (2002). A Framework to Evaluate the Informatization Level. In W. Van Grembergen (Ed.), Information Systems Evaluation Management, pp. 287-298, IGI Global Scientific Publishing: https://doi.org/10.4018/978-1-931777-18-6.ch018.

Lundberg, S.M., and Lee, S.-I, (2017), A Unified Approach to Interpreting Model Predictions, In Proceedings of the Advances in Neural Information Processing Systems, 4765-4774.

Moon, D., Yoon, S., Choi, S., and Kim H., (2020), A Machine Learning Approach for the Success Prediction of Reward Crowdfunding Project. Korea Business Review, 24(3), 125-143. 10.17287/kbr.2020.24.3.125. **

Natekin, A., and Knoll, A., (2013), Gradient boosting machines, a tutorial, Frontiers in Neurorobotics, 7, 21, 1-21.

Oh, J., Lee, Y., and Kim, G., (2020), Improvement of Solar Power Forecasting Using Interpretation of Artificial Intelligence. The transactions of The Korean Institute of Electrical Engineers, 69(7), 1111-1116. 10.5370/KIEE.2020.69.7.1111. **

Park, D., and Kwahk, K., (2020), The Effects of Information Systems Based Working Environment on the Performance of SMEs. Korean Management Review, 49(1), 215-249. **

Peng, J., Hahn, J., and Huang, K.-W, (2022), Handling Missing Values in Information Systems Research: A Review of Methods and Assumptions, Information Systems Research, 34, 1, 5-26.

Roberts, N., and Grover, V., (2012), Leveraging Information Technology Infrastructure to Facilitate a Firm's Customer Agility and Competitive Activity: An Empirical Investigation, Journal of Management Information Systems, 28, 4, 231-269.

Rosemann, M., and vom Brocke, J., (2010), The Six Core Elements of Business Process Management, Handbook on Business Process Management 1, Springer, Berlin.

Song, M.-K., Lin, F.-C., Ward, S.E., and Fine, J.P., (2013), Composite Variables: When and How, Nursing Research, 62, 1, 45-49.

Sun, J., (2022), The Impacts of IS(Information Systems) Use on Firm Performance, 45(3), 215-229. http://dx.doi.org/10.22778/jci.2022.45.3.215. **

Thiruvengadam, K., Watson, B., Ponnuraja, C., and Rajendran, K., (2022), A Review of Statistical Modelling and Machine Learning in Analytical Problems, International Journal of Applied Engineering Research, 17, 5, 506-510.

Van Looy, A., and Shafagatova, A., (2016), Business process performance measurement: a structured literature review of indicators, measures and metrics, SpringerPlus, 5, 1, 1-24.

van Zyl, W.R., Henning, S., and van der Poll, J.A., (2022), A Framework for Knowledge Management System Adoption in Small and Medium Enterprises, Computers, 11, 9, 128-151.

Viktor, K., Alexey, K., Andrey, D., and Alexander, F., (2020), Assessing the Enterprise Informatization Level in Digital Economy Conditions, IOP Conference Series: Materials Science and Engineering, 940, 012012.

Wang, Z., and Wang, N., (2012), Knowledge sharing, innovation and firm performance, Expert Systems with Applications, 39, 10, 8899-8908.

Won, J., and Lee, K., (2021), Analysis of the Informatization Factors of Small and Medium Enterprises Using the IT Business Value Model. Information Systems Review, 23(1), 135-154. **

Yi, C., (2021), The Prediction Power of Busan's Strategic Industry on the Income through Ensemble Machine Learning Model. Journal of Economics Studies, 39(4), 3-29. **

Yoon, Y., Kim T., and Kim, S., (2022). Study on Predicting the Designation of Administrative Issue in the KOSDAQ Market Based on Machine Learning Based on Financial Data. Asia-Pacific Journal of Business Venturing and Entrepreneurship, 17(1), 229-249. **

** : Papers written in Korean