

## Evaluating Digital Transformation and Low-Carbon Readiness in Taiwanese Industries: A Machine Learning Approach



Christine Dewi<sup>1</sup>, Abbot Po Shun Chen<sup>2\*</sup>, Rio Arya Andika<sup>1</sup>, Kartika Gianina Tileng<sup>3</sup>, Stephen Aprius Sutresno<sup>4</sup>, Dalianus Riantama<sup>5</sup>, Chang Hsiung Chen<sup>6</sup>, Ya Ying Lai<sup>6</sup>

<sup>1</sup> Department of Information Technology, Satya Wacana Christian University, Salatiga 50711, Indonesia

<sup>2</sup> Department of Marketing and Logistics Management, Chaoyang University of Technology, Taichung City 413310, Taiwan

<sup>3</sup> Department of Information Systems, School of Information Technology, Universitas Ciputra, Surabaya 60219, Indonesia

<sup>4</sup> Information System Department, School of Bioscience, Technology, and Innovation, Atma Jaya Catholic University of Indonesia, Jakarta 12930, Indonesia

<sup>5</sup> Management Department, BINUS Business School Undergraduate Program, Bina Nusantara University, Jakarta 11480, Indonesia

<sup>6</sup> Process Efficiency Improvement Department, Green Technology Division, Central Region Campus, Industrial Technology Research Institute (ITRI), Nantou 540219, Taiwan

Corresponding Author Email: [chprosen@gm.cyut.edu.tw](mailto:chprosen@gm.cyut.edu.tw)

Copyright: ©2026 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/isi.310223>

### ABSTRACT

**Received:** 23 November 2025

**Revised:** 20 January 2026

**Accepted:** 18 February 2026

**Available online:** 28 February 2026

#### **Keywords:**

*machine learning, digital transformation readiness, low-carbon transition, industrial sustainability, policy analysis*

Taiwan's manufacturing sectors are undergoing transformation amid global digitalization and low-carbon pressures. This study develops a machine learning (ML) framework to assess industrial readiness for digital and low-carbon transitions. By integrating spatial data from land use and agricultural clusters, six models—Decision Tree, XGBoost, Random Forest, LightGBM, Logistic Regression (LR), and SVM—were tested using features such as digital infrastructure, policy environment, workforce capacity, and access to clean energy. SVM and LR achieved the highest recall in identifying industries prepared for transition, while XGBoost and LightGBM provided an effective balance between precision and recall. The proposed framework offers data-driven insights to guide policymakers in supporting industrial modernization and sustainability across regions. It emphasizes leveraging ML to analyze complex industrial data, helping to predict the readiness of various sectors to adopt digital transformation and low-carbon practices. By providing actionable insights, this study offers a scalable and adaptable model for evaluating readiness, contributing to industrial policy development and helping to align industrial growth with global sustainability goals. This approach is valuable for improving regional development strategies, ensuring that industrial sectors are better equipped for technological and environmental changes. The study's findings advance the application of ML in industrial policy planning, offering practical solutions for industries undergoing digital and low-carbon transformations.

## 1. INTRODUCTION

The digital transformation and the implementation of low-carbon industries have become two key agendas in the development of industrial zones across various countries, including Taiwan [1, 2]. With increasingly stringent environmental regulations and rising demands for production efficiency, companies within industrial zones must adapt through technology-driven innovation [3, 4]. One strategic approach to support this shift is through land use planning, integrated with the development of agricultural industry cluster areas [5, 6]. Studies on land use planning and industrial clusters provide an essential foundation for understanding the dynamics of digital transformation readiness and the adoption of low-carbon principles at the regional level. In the context of Taiwan, leveraging land use planning can help optimize

location, resources, and infrastructure to accelerate industrial adaptation [7, 8].

However, the readiness of each industrial zone to undergo transformation is not uniform and depends on a range of interconnected factors [9–11]. Therefore, an accurate prediction method is required to assess the readiness for digital transformation and the implementation of low-carbon industry principles. Machine Learning (ML) presents significant opportunities for analyzing this complex data, owing to its ability to uncover hidden patterns and make predictions based on big data. In relation to land use planning and industrial clusters, ML enables the identification of key features that serve as indicators of industrial zone readiness. This study aims to develop a predictive model for assessing the readiness of industrial zones in Taiwan for transformation by integrating ML techniques with regional planning concepts [12, 13].

Identifying the most influential determinants is essential in developing an efficient and reliable predictive model. Preparedness of industrial areas is influenced by numerous determinants like digital infrastructure, policy and regulatory environments, workforce characteristics, and clean energy accessibility thus it is indispensable to identify such key determinants to achieve the best possible analytical results. In addition to enhancing the accuracy of the model, this phase also provides policymakers and developers with transparent directions [14, 15]. Focusing on such key elements, the study strives to unveil the main driving forces of digitalization and sustainability. This approach follows the fundamental principle of land use planning in which data-centric policies are emphasized to promote better regional and industrial development [16, 17].

This research is purposes to integrate results from land use planning [18, 19], agricultural industry clusters [20, 21], and ML algorithms to provide theoretically relevant and practically useful contributions to industrial area development in Taiwan. In addition, the outcomes of the research are expected to serve as a policymaking framework which supports industrialization with sustainability and technical progress in the coming times. The predictive models and feature selection methods constructed in this study are expected to apply to other countries facing similar challenges in promoting digitalization and supporting a low-carbon economy. Overall, the research aims to show how data analytics and innovation are best used to assist industrial sectors to achieve progress through better and more sustainable planning methods.

The main contribution of this research lies in the development of an integrated ML-based predictive framework. This framework combines land use planning, agricultural industry cluster analysis, and key industrial indicators. It is designed to assess the readiness of Taiwan's industrial zones for digital transformation and low-carbon development. Unlike previous studies that treat digitalization and sustainability separately, this research presents a unified, data-driven approach that accounts for spatial, infrastructural, and socioeconomic variables across regions. By evaluating and comparing the performance of various ML models, the study identifies the most effective techniques for classifying industrial readiness, while also uncovering the most influential determinants such as digital infrastructure, clean energy access, workforce characteristics, and policy environments [22-24]. The proposed framework not only improves predictive accuracy but also provides practical insights for policymakers, urban planners, and industry stakeholders, enabling them to make targeted decisions to support sustainable industrial growth. Additionally, the methodology offers a scalable and transferable model that can be adapted for use in other countries facing similar industrial modernization and climate transition challenges.

This research is important because it provides a data-driven framework for understanding how well-prepared Taiwan's industrial sectors are for two of the most critical transitions of our time: digital transformation and low-carbon development. By leveraging ML techniques and integrating spatial and sectoral data, the study offers a nuanced and scalable method for identifying readiness gaps across industries and regions. This approach supports evidence-based policymaking, enabling government and industry leaders to allocate resources more effectively, prioritize strategic interventions, and design targeted support mechanisms. Furthermore, the dual focus on

digitalization and sustainability reflects the real-world challenges industries face in adapting to technological change while meeting environmental goals. The methodology is not only applicable to Taiwan but also offers a transferable model for other countries seeking to align industrial modernization with global sustainability agendas [25].

Digital transformation and low-carbon transition can be viewed as part of broader industrial transformation processes driven by technological innovation and sustainability pressures. According to innovation theory, firms adopt new technologies to improve productivity, efficiency, and competitiveness. In the context of industrial policy, digital technologies and low-carbon solutions represent key innovation pathways that enable industries to adapt to changing regulatory environments and global sustainability requirements. Despite the increasing attention to digital transformation and low-carbon development in industrial sectors, empirical studies that systematically evaluate industrial readiness using data-driven approaches remain limited. Most existing research focuses on qualitative policy analysis or sector-specific case studies, leaving a gap in quantitative frameworks capable of identifying key determinants of industrial readiness.

Furthermore, the integration of ML techniques with industrial survey data to assess digital transformation and low-carbon readiness across sectors has not been widely explored. Addressing this gap, this study applies multiple ML models and feature selection methods to analyses industrial readiness in Taiwanese industries.

## **2. LITERATURE REVIEW**

### **2.1 Digital transformation in industrial development**

Digital transformation has become a key driver of industrial modernization and productivity improvement in the era of Industry 4.0. The adoption of digital technologies such as the Internet of Things (IoT), artificial intelligence (AI), and cloud computing enables firms to optimize production processes, improve operational efficiency, and support data-driven decision making [26-29]. These technologies facilitate real-time monitoring, predictive maintenance, and resource optimization, which can significantly enhance industrial competitiveness.

Previous studies have also highlighted the role of digital capabilities in enabling sustainable manufacturing practices. For instance, digital technologies allow firms to track energy consumption and optimize resource allocation, contributing to reductions in carbon emissions and improved environmental performance. As a result, digital transformation is increasingly viewed as a critical enabler of sustainable industrial development. However, existing research often focuses on the technological adoption of individual firms or specific sectors. Less attention has been given to evaluating digital transformation readiness at the regional or industrial cluster level, particularly using data-driven analytical approaches.

### **2.2 Low-carbon industrial transition**

In parallel with digital transformation, the transition toward low-carbon industrial systems has become a global policy priority. Governments and international organizations increasingly promote carbon reduction strategies through

environmental regulation, renewable energy integration, and green innovation initiatives [30].

Research on low-carbon industrial development emphasizes the importance of technological innovation, environmental policy frameworks, and industrial restructuring. Several studies have examined the role of green technologies and renewable energy adoption in reducing industrial emissions. In addition, regional development studies have highlighted the importance of spatial planning and industrial clustering in supporting sustainable development.

Industrial clusters can promote knowledge spillovers, resource sharing, and infrastructure efficiency, which are critical factors in achieving both economic and environmental objectives [31-35]. Despite these insights, few studies provide a comprehensive framework for evaluating the readiness of industrial regions to transition toward low-carbon development.

### 2.3 Machine learning for industrial readiness assessment

The rapid growth of ML has created new opportunities for data-driven analysis in policy-relevant domains. ML techniques have been widely applied in areas such as smart city planning, energy demand forecasting, and economic resilience modeling [36].

Compared with traditional statistical approaches, ML algorithms can effectively model complex nonlinear relationships and handle high-dimensional datasets. As a result, they have increasingly been used to analyze industrial performance, sustainability indicators, and economic development patterns [37-39]

Models such as Random Forest, Support Vector Machine, and gradient boosting algorithms have demonstrated strong predictive performance in classification and prediction tasks involving multidimensional data. However, many studies apply these techniques to narrow research problems, such as energy consumption forecasting or firm-level performance analysis, rather than assessing broader industrial transformation readiness.

### 2.4 Research gap

Although previous studies have examined digital transformation, low-carbon development, and ML applications separately, limited research integrates these perspectives into a unified analytical framework.

In particular, few studies apply ML techniques to evaluate industrial readiness for both digital transformation and low-carbon transition simultaneously. Furthermore, existing research rarely incorporates spatial and sectoral planning variables, such as land use characteristics and industrial clustering patterns, into predictive models of industrial transformation.

To address this gap, this study applies multiple ML models to industrial survey data from Taiwan in order to evaluate readiness for digital transformation and low-carbon development. By integrating spatial planning data with ML-based classification techniques, this research provides a comprehensive framework for assessing industrial transformation at the regional level.

### 2.5 Research contributions

To address the research gap identified in the previous section, this study provides several theoretical and practical

contributions to the field of industrial transformation and sustainability assessment.

Understanding the readiness of industrial zones for digital transformation and low-carbon development is essential for shaping effective regional policies and sustainable growth strategies. This study integrates ML techniques with spatial and sectoral planning concepts to evaluate industrial readiness using a data-driven framework. The proposed approach not only improves predictive accuracy but also provides actionable insights that are relevant to policymakers, industry stakeholders, and urban planners.

Table 1 summarizes the main contributions and practical benefits of the proposed framework in supporting Taiwan's broader goals for industrial modernization and environmental sustainability.

**Table 1.** Summary of research contributions

| No. | Benefit  | Key   |
|-----|--|---|
| 1   | Evidence-Based Industry Assessment [40, 41]    | Enables data-driven evaluation of industrial readiness, moving beyond generalized policy assumptions to detailed regional and sectoral analysis.                              |
| 2   | Application for Machine Learning (ML) [42, 43] | Utilizes advanced ML algorithms (e.g., Support Vector Machine (SVM), XGBoost) for high-accuracy predictions and identification of readiness patterns across complex datasets. |
| 3   | Dual Focus: Digital and Low-Carbon [44, 45]    | Simultaneously assesses both digital transformation and sustainability, addressing two key priorities in industrial development.  |
| 4   | Spatial and Sectoral Integration [46]          | Incorporates land use and industry cluster data to capture geographic and structural differences, enhancing relevance for regional planning.                                  |
| 5   | Strategic Policy Guidance [47]                 | Provides insights for targeted interventions, investment prioritization, and workforce or infrastructure development in underperforming zones.                                |
| 6   | Transferable Framework [48]                    | Offers a replicable methodology that can be applied in other countries facing similar challenges in industrial modernization and climate alignment.                           |

## 3. METHOD

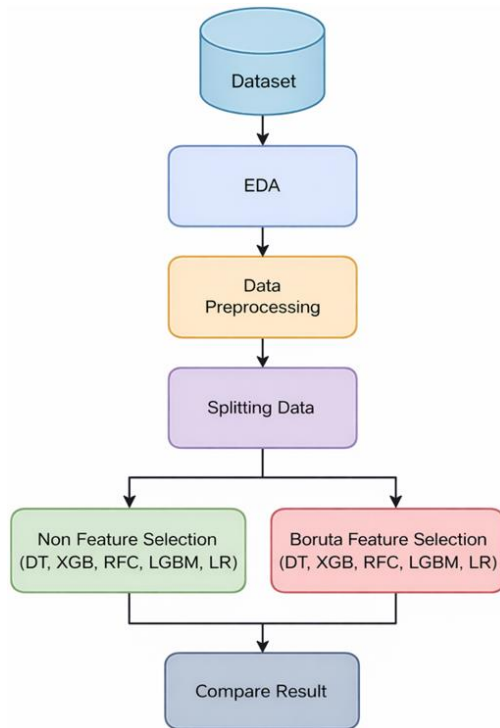
### 3.1 Research design

The objective of the research is to build a predictive model to evaluate the Taiwanese industrial sectors' readiness to become digitally and low-carbon-compliant using several ML techniques. As shown in Figure 1 the data will also undergo Exploratory Data Analysis (EDA) first to check data distribution, identify missing values, and look for potential outliers on data quality. Both descriptive statistics and data visualization will also be utilized to build a richer understanding of the nature of the dataset.

Following EDA, the next phase is data preprocessing that includes missing value handling, encoding of categorical variables and standardization or normalizing numerical variables. The data to be used will be further processed to preserve its quality and suitability to proceed to the Data Split phase. The data will be divided into a train and a test set in a standard split of 80:20. The split will permit the model to learn

on the training data and check its performance on fresh, unseen data.

The pre-processed data were subsequently used to train several ML algorithms, including Decision Tree (DT), XGBoost (XGB), Random Forest Classifier (RFC), Light Gradient Boosting Machine (LGBM), and Logistic Regression (LR). All models were implemented in Python using widely adopted ML libraries, including scikit-learn, XGBoost, and LightGBM. The experiments were conducted using Python 3.12 in Google Colab, with scikit-learn for Decision Tree, Random Forest, and LR, the xgboost package for XGBoost, and the lightgbm library for LightGBM. Unless otherwise specified, all algorithms were trained using the default hyperparameters provided by their respective libraries to ensure consistent and fair comparison across models.



**Figure 1.** Research workflow

To investigate the impact of feature selection on model performance, two experimental settings were conducted. In the first setting, the models were trained using the complete set of features. In the second setting, feature selection was applied using the Boruta algorithm, implemented through the BorutaPy library in Python, which is built on top of the Random Forest estimator from scikit-learn. Boruta was used to identify the most relevant variables. The same ML models were then retrained using the selected features, allowing a direct comparison between models trained with all features and those trained with the Boruta-selected subset.

Prior to model training, the dataset was examined for missing values and potential outliers. Records containing missing values were removed to ensure data completeness and maintain consistency across all features used in the analysis. Since the proportion of missing values was 0, removing these records did not significantly affect the dataset representation.

Potential outliers were inspected during the exploratory data analysis phase using statistical summaries and visualization techniques such as boxplots and distribution plots. Because most variables originate from structured survey responses with predefined categorical options, extreme outliers were minimal

and no aggressive outlier removal was required.

### 3.2 Dataset description

The dataset used in this study was derived from a structured survey conducted collaboratively with the Industrial Technology Research Institute. The survey aimed to collect comprehensive data reflecting the readiness of industrial sectors in Taiwan to undertake digital transformation and transition toward low-carbon practices. The survey was administered through a structured questionnaire distributed to firms registered within relevant industrial categories. Firms were selected from registered industrial sectors that are relevant to digital transformation and low-carbon transition. It covered multiple manufacturing sectors, including electronics, machinery, metal processing, and chemical manufacturing. To ensure adequate representation, the survey targeted firms located in major industrial regions across Taiwan, encompassing both northern and southern industrial clusters.

**Table 2.** Summary of dataset variables

| Category                            | Description   | Example Variables   |
|-------------------------------------|---|---|
| Firm Characteristics                | Basic information describing firm profile and operational scale                     | capital amount (h), factory area (i), number of employees (l), product type (g)           |
| Location and Land Status            | Variables describing land ownership and zoning conditions                           | land use zone (d), land use type (e), land ownership (n), rent/share/state land (n11–n14) |
| Industrial Sector                   | Variables describing industrial classification and sectoral identity                | industry type (f)   |
| Labor Resources                     | Indicators related to migrant worker demand and labour requirements                 | migrant workers need (m1), hired migrant workers (m11), additional workers required (m12) |
| Land Development Planning           | Firm readiness to apply for land-use planning programs                              | willingness to apply for land plan (o), application status (o1–o4)                        |
| Technological Readiness             | Indicators related to digital transformation and production technology guidance     | intelligent technology guidance (p1), low-carbon technology guidance (p2)                 |
| Energy and Environmental Transition | Variables related to renewable energy adoption and sustainability practices         | solar photovoltaic installation (q), solar planning status (q1–q4)                        |
| Environmental Compliance            | Indicators related to wastewater management and environmental regulation compliance | wastewater discharge permit (t), discharge method (t11–t15)                               |
| Environmental Infrastructure        | Environmental mitigation measures such as green belts                               | isolation green belt facilities (r)   |
| Financial Capacity and Compliance   | Ability of firms to pay environmental contributions or fees                         | monetary contribution payment (s)   |

The primary objective of the survey was to gather

information related to industrial operational characteristics, technological readiness, and environmental transition capacity, particularly in relation to digital transformation and low-carbon initiatives.

The final dataset comprises 5,000 records, with each record representing an individual industrial entity. A total of 97 features were collected, capturing critical dimensions that influence industrial readiness. These dimensions include digital infrastructure, energy usage patterns, workforce competency, regulatory environment, accessibility to clean energy, and organizational willingness to pursue transformation. Together, these features reflect both internal and external factors that shape an industry's capacity to adapt to technological and environmental changes.

| 二、生產製程技術需求   | 智能化  | 低碳化                         | 未來有需求                       | 未來無需求                                  |
|--|--|-----------------------------|-----------------------------|--|
| 請勾選一項最需要的<br>Please select one                             | <input type="checkbox"/> p1                                | <input type="checkbox"/> p2 | <input type="checkbox"/> p3 | <input checked="" type="checkbox"/> p4 |
| 選擇「未來有需求」，續勾選以下選項：<br>If you choose p3 please select (可複選) | 選擇「未來無需求」，續勾選以下選項：<br>If you choose p4 please select (可複選) |                             |                             |  |
| p11 <input type="checkbox"/> 想瞭解智能化技術輔導相關內容                | p21 <input type="checkbox"/> 不清楚輔導內容或效益及申請管道               |                             |                             |  |
| p12 <input type="checkbox"/> 想瞭解低碳化技術輔導相關內容                | p22 <input type="checkbox"/> 公司目前營運不穩定                     |                             |                             |  |
| p13 <input type="checkbox"/> 想瞭解低碳化或智能化轉型貸款                | p23 <input checked="" type="checkbox"/> 等公司未來規模擴大再說        |                             |                             |  |
| p14 <input type="checkbox"/> 其他:                           | p24 <input type="checkbox"/> 其他:                           |                             |                             |  |

**Figure 2.** Questionnaire example related to p1: 'Smart Technology Guidance' dan p2: 'Low Carbon Guidance'

Figure 2 shows the Questionnaire example related to p1: 'Smart Technology Guidance' dan p2: 'Low Carbon Guidance'. Table 2 presents the explanations of the questions related to P1 and P2. Table 3 presents the explanation of the questions related to P1 and P2, which were designed to capture the key dimensions of the study variables.

This survey section focuses on production process technology needs and begins by asking respondents to select the single most important option from four categories: p1 (Intelligent Technology), p2 (Low-Carbon Technology), p3 (Future Demand), and p4 (No Future Demand). If a respondent chooses p3 (Future Demand), they may select multiple sub-options, including: p11 (wanting to learn more about intelligent technology consulting), p12 (wanting to learn more about low-carbon technology consulting), p13 (wanting to learn about low-carbon or intelligent benchmarking data), and p14 (other). If a respondent chooses p4 (No Future Demand), they may also select multiple sub-options, including: p21 (unclear about the guidance content, benefits, or application/consultation channels), p22 (company's current operations are unstable), p23 (will reconsider when the company expands in the future), and p24 (other). In the example shown, the respondent selected p4 (No Future Demand) and, within this category, marked p23, indicating that they will reconsider these technologies when the company expands in the future.

### 3.3 Exploratory data analysis

As an initial step, this research conducted a feature correlation analysis to gain deeper insights into the interrelationships among variables within the dataset. The results, illustrated in Figure 3, indicate a strong interest among many industries in transitioning toward low-carbon products.

From the heatmap shown in Figure 4, it is evident that certain features exhibit strong positive correlations. For instance, p3 is highly correlated with p11 (0.75) and p12 (0.78), indicating that these variables may capture similar underlying

patterns in the dataset. Similarly, a strong positive correlation is observed between p4 and p23 (0.72). Such relationships suggest that several features may contain overlapping information. In contrast, strong negative correlations are also observed, such as between p3 and p4 (-0.63) and between p4 and p12 (-0.49), indicating that increases in one variable tend to be associated with decreases in the other.

**Table 3.** Question explanation related to P1 and P2

| Code | Chinese         | English  |
|------|-----------------|--|
| p    | 生產製程技術輔導        | Production process technical guidance  |
| p1   | 智能化             | Intelligent  |
| p2   | 低碳化             | Low carbonization  |
| p3   | 未來有需求           | There will be demand in the future   |
| p4   | 未來無需求           | No demand in the future  |
| p1   | 未來有需求智慧低碳       | There will be demand for smart low carbon in the future  |
| p11  | 想瞭解智能化技術輔導相關內容  | I want to learn more about smart technology guidance.  |
| p12  | 想瞭解低碳化技術輔導相關內容  | I want to learn more about low-carbon technology guidance  |
| p13  | 想瞭解低碳化或智能化轉型貸款  | I want to know about low-carbon or smart transformation loans  |
| p14  | 其它              | other  |
| p2   | 未來無需求智慧低碳       | There will be no demand for smart low-carbon solutions in the future.                                  |
| p21  | 不清楚輔導內容或效益及申請管道 | Unclear about the content or benefits of the guidance and the channels for application or consultation |
| p22  | 公司目前營運不穩定       | The company's current operations are unstable  |
| p23  | 等公司未來規模擴大再說     | We will discuss it when the company expands in the future  |
| P24  | 其它              | other  |

Several features, including h, i, and l, as shown in Figure 5, demonstrate very weak correlations with most other variables. This suggests that these features may contain relatively

independent information that could be beneficial for the model, as they are less likely to introduce redundancy. Similarly, feature p24 shows consistently weak correlations with other predictors, indicating minimal overlap with the remaining variables and potential usefulness in the learning process.

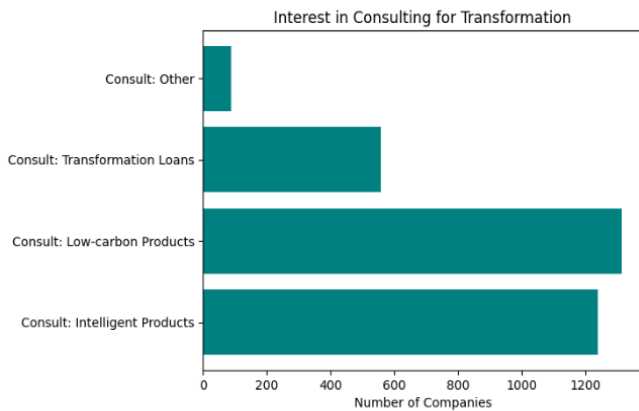


Figure 3. Transformation interest

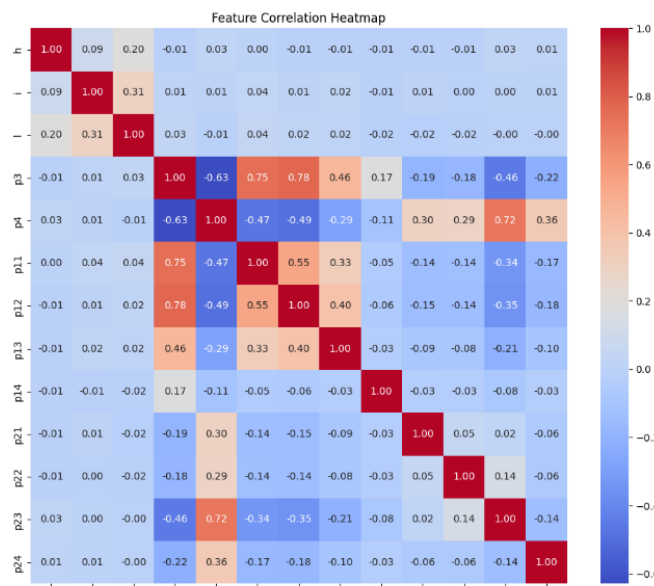


Figure 4. Heatmap correlation



Figure 5. Industries capacity factors

Correlation analysis was performed to identify multicollinearity and interdependencies among predictors. Strong positive correlations were observed between digital infrastructure and workforce skill level, indicating that regions with higher digital access tend to have a more technologically capable labor force.

However, several features—such as clean energy access and policy support—exhibited weak correlations with digital factors, suggesting the need for a multidimensional modeling approach. Additionally, boxplots and histograms were used to examine outliers and class distribution. The readiness labels exhibited a class imbalance, with most industrial zones falling into the moderate readiness category, and fewer zones classified as highly ready or severely underprepared. This justified the later emphasis on models and evaluation metrics (e.g., recall, F1-score) that are robust to class imbalance.

### 3.4 Feature selection with Boruta

Boruta is a feature selection method based on the random forest classification approach. It operates by generating shadow features, which are shuffled copies of the original variables, and then fitting a random forest classifier to the extended dataset. The method repeatedly evaluates the importance of each original feature by comparing its score with those of the shadow features. Features that achieve higher importance than their shadow counterparts are considered relevant and kept, whereas features with lower importance are regarded as irrelevant and discarded [49].

In this study, Boruta was employed as a preprocessing step to identify key determinants that influence the readiness of industrial zones for digital transformation and low-carbon development. Given the multidimensional nature of the dataset—which includes variables related to digital infrastructure, workforce skills, regulatory support, spatial planning, and clean energy access, Boruta was particularly effective in filtering out noise and reducing dimensionality without sacrificing important information. By selecting the most influential features, Boruta not only improved the computational efficiency of the downstream ML models but also strengthened the policy relevance of the analysis. The selected features provided clearer insights into the primary drivers of industrial readiness, enabling more targeted recommendations for regional development and sustainability planning.

Compared to traditional feature selection methods, Boruta offers the advantage of being model-agnostic within the tree-based family and maintains high sensitivity to weak but important predictors, making it especially useful in complex socio-technical datasets such as those used in this study. The integration of Boruta contributed to enhanced model performance and interpretability, forming a critical component of the research methodology.

### 3.5 Machine learning models

#### 3.5.1 Decision Tree

DT is one of the most widely used supervised learning algorithms for classification tasks due to its interpretability, simplicity, and ability to handle both categorical and numerical data [50]. It works by recursively splitting the dataset based on the feature that provides the highest information gain or lowest Gini impurity, forming a tree-like structure where each internal node represents a test on a

feature, each branch represents the outcome of the test, and each leaf node represents a class label. A Decision Tree is a class of ML approaches known by its hierarchical tree structure and used in both classification and regression tasks. The algorithm operates by recursively splitting the data into ever-smaller partitions bearing on the values in different attributes using impurity measurements like the Gini Index or Entropy [51]. To the ID3 algorithm, the best attribute to select is found by the calculation of Information Gain (IG), which is defined in Eq. (1):

$$IG(D, A) = H(D) - \sum_{v \in A} \frac{|D_v|}{|D|} H(D_v) \quad (1)$$

In the context of land use planning and industrial zone analysis, a Decision Tree is employed to map the readiness of areas based on spatial and economic attributes. While a Decision Tree is easy to interpret, it is prone to overfitting, particularly when dealing with large and complex datasets [52].

In the context of mapping digital transformation and low-carbon readiness in Taiwanese industries, the Decision Tree model serves as a baseline classifier to evaluate the influence of individual factors such as digital infrastructure, workforce characteristics, policy environment, and clean energy access. Its hierarchical structure is particularly useful for understanding how these features interact to determine industrial readiness. In this study, the Decision Tree model was used to classify industrial zones based on their readiness levels. Its performance was compared to more complex ensemble and linear models to evaluate trade-offs between interpretability and predictive power. The results indicate that while Decision Tree provided reasonable classification results, its recall for minority classes was relatively lower, suggesting that it may be better suited for exploratory analysis or as a component in an ensemble framework.

### 3.5.2 XGBoost

XGBoost (Extreme Gradient Boosting) is a very potent tree-based boosting algorithm aiming to bring a dramatic boost in the predictive ability of the models with a high level of computing efficiency. XGBoost reduces the objective function by reducing prediction errors using robust gradient descent optimization algorithms. XGBoost also uses regularization to avoid models to overfit, i.e., when models are overly tailored to the data on which it has been trained and do not work on other data. Not only does it enhance the model's performance, but it also maintains it stable if the model is extended to novel, unseen data [53-55]. XGBoost objective function used to achieve above is defined in Eq. (2):

$$Obj(\theta) = \sum_{i=1}^n l(\hat{y}_i, y_i) + \sum_{k=1}^K \Omega(f_k) \quad (2)$$

With regularization, as shown in Eq. (3):

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda |w|^2 \quad (3)$$

Kwaghtyo et al. [54] used XGBoost to predict smart farming prediction models for precision agriculture with a more than 99% accuracy. XGBoost possesses outstanding ability to deal with very large datasets, missing values, and correlation between features. These characteristics make it a

viable method to predict the readiness of low-carbon and smart industrial parks.

### 3.5.3 Random Forest

Random Forest is a robust ensemble learning technique that builds a large forest of decision trees using a technique called bootstrap aggregating or bagging. There, a randomly sampled subset of the data is used to train each tree to add diversity to the predictions of the trees. The overall prediction of the model is computed by combining the prediction of all the trees together. In classification, it is done through majority vote, i.e., the mode class receiving the maximum votes by the trees is the overall prediction of the model. In the case of regression, the prediction is done by computing the meaning of all the predictions of the trees [46, 47]. The mathematical representation of the Random Forest predictions in the case of classification is defined in Eq. (4):

$$\hat{y} = mode\{h_1(x), h_2(x), \dots, h_K(x)\} \quad (4)$$

where,  $h_k$  represents the prediction from the k-th tree [56]. Yang et al. [57] employed Random Forest to analyze The Impact of Smart Manufacturing Demonstration Projects on Green Innovation. With its strength in reducing overfitting and handling high-dimensional data, Random Forest has become a go-to method in studies that predicts outcomes based on the characteristics of industrial zones [58-60].

### 3.5.4 Light Gradient Boosting Machine

LGBM is a boosted algorithm optimized to perform fast and efficient bootstrapping with histogram-based learning methods. Unlike XGBoost which does level-wise growing of trees, LGBM does leaf-wise growing of trees and is thus more aggressive in terms of finding optimal splits [49-51]. The loss function optimized by LGBM is formulated as defined in Eq. (5):

$$L = \sum_{i=1}^n l(y_i, \hat{y}_i) + \lambda \sum_{j=1}^d w_j^2 \quad (5)$$

Benos et al. [61] used LGBM to predict the uptake of blockchain in agriculture with higher speed and precision than existing models. The ability of LGBM to handle massive data makes it very relevant to assess industrial area readiness to attain digitalization and low-carbon transition.

In this study, LightGBM was employed to classify the readiness of Taiwanese industrial zones for digital transformation and low-carbon development. Its ability to handle large feature spaces and manage categorical variables effectively made it well-suited for modeling complex relationships among diverse input features, including digital infrastructure, policy incentives, workforce capabilities, and clean energy access. Compared to other ensemble methods like Random Forest and XGBoost, LightGBM demonstrated a superior balance between precision, recall, and F1-score, particularly in scenarios with class imbalance. Its leaf-wise splitting strategy often results in deeper trees that capture intricate patterns in the data, enabling it to perform well in distinguishing high- and low-readiness zones—even when such cases are underrepresented. Moreover, LightGBM offers built-in support for feature importance ranking, which enhances model interpretability and aids in identifying key drivers of transformation readiness. These insights are

particularly valuable for policymakers and industrial planners aiming to target interventions effectively. Overall, LightGBM contributed robust predictive performance and computational efficiency in this research, making it a practical choice for readiness classification at a national scale.

### 3.5.5 Logistic Regression

LR is a widespread statistical model used to predict the probability of a two-outcome event as a function of a single or multiple predictor variable. The model uses the logit function to define a relationship between the predictor variables and the probability of the event occurring successfully. The model also ensures the returned probability result to fall between a range of 0 to 1 and hence is best used in a binary-classification problem [62]. The general LR formula is as defined in Eq. (6):

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p)}} \quad (6)$$

Lin et al. [63] used LR to describe Taiwanese small and medium-sized manufacturing firms' readiness towards digitalization. LR remains a popular standard model because of the straightforward interpretability of its coefficients and by which the impact of a single predictor is readily understandable. It is, however, less capable of modeling more complex and non-linear relations and as such its performance might be restricted to more complex datasets.

## 3.6 Performance evaluation

To ensure the reliability and effectiveness of the proposed predictive models, a comprehensive performance evaluation was conducted using standard classification metrics. The models—including Decision Tree, Random Forest, XGBoost, LightGBM, LR, and Support Vector Machine—were assessed based on their ability to classify industrial zones according to their readiness for digital transformation and low-carbon development.

Key performance metrics used in this study include accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the model but may be misleading in the presence of class imbalance. Precision quantifies the proportion of true positives among all predicted positives, highlighting the model's ability to avoid false alarms. Recall, or sensitivity, indicates the model's ability to correctly identify all relevant instances, particularly important for detecting underrepresented readiness classes. F1-score provides a harmonic means of precision and recall, offering a balanced measure that is especially useful when the dataset is skewed.

Given that industrial readiness levels were not uniformly distributed, some zones were significantly more prepared than others, class imbalance was a critical consideration. Therefore, models with high recall and F1-score for minority classes were prioritized, as they are more effective in identifying sectors that may require urgent policy intervention or support. In addition to these metrics, confusion matrices were generated to visually assess model performance across readiness classes. The comparison revealed that ensemble models such as XGBoost and LightGBM achieved the most balanced results across all metrics. SVM and LR were particularly effective in achieving high recall, making them valuable for detecting highly ready or unprepared zones. Meanwhile, Decision Tree, although interpretable and fast, showed lower recall for minority classes, limiting its standalone application in critical

readiness assessments. This evaluation framework ensures that model selection is aligned not only with statistical performance but also with practical policy objectives, such as maximizing the detection of vulnerable or high-potential industrial zones.

The formula for accuracy, i.e. the percentage of correct predictions out of the overall total, is defined in Eq. (7):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

where: TP: True Positive, TN: True Negative, FP: False Positive, and FN: False Negative.

Besides that, precision and recall also show how good the model is at the positive class. Precision measures how good the positive predictions are, and recall provides how good the model is at identifying the positive class. All of the above metrics can be computed from the following equations which is defined in Eqs. (8) and (9):

$$\text{Precision} = \frac{TP}{TP + FP} \quad (8)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (9)$$

The F1-score is the harmonic means of recall and precision and provides a balance between the two. The F1-score is very useful when balancing the trade-off between recall and precision. The F1-score is calculated as defined in Eq. (10):

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

Furthermore, this study also utilizes Receiver Operating Characteristic (ROC) Curve and Precision-Recall (PR) Curve to enhance performance evaluation, particularly under class imbalance. The ROC Curve plots the True Positive Rate (Recall) against the False Positive Rate, and the Area Under the Curve (ROC AUC) represents the model's ability to distinguish between classes. A higher AUC indicates better discrimination performance. However, in scenarios with a high-class imbalance, PR Curve is often more informative, as it focuses on the performance with respect to the positive (minority) class. The area under the PR Curve provides a measure of how well the model identifies positive instances without being misled by the abundance of negative instances [64].

These additional metrics ensure a more robust and comprehensive assessment of model performance beyond conventional accuracy, particularly when the goal is to minimize false negatives or optimize detection of the ready industries.

## 3.7 Stratified K-Fold

Stratified K-Fold is a cross-validation technique that ensures each fold used in training and testing preserves the same proportion of class labels as the original dataset. Unlike regular K-Fold, which can produce imbalanced splits especially in classification tasks with skewed class distributions, Stratified K-Fold is designed to maintain the class balance across all folds. This is particularly important for classification problems involving minority classes, such as

predicting readiness in this study, where the positive class (readiness = 1) is underrepresented.

In each iteration, the data is split into k equal-sized folds, and the model is trained on k-1 fold and validated on the remaining fold. This process is repeated k times, with each fold being used once as the validation set. After all iterations, the predictions from each fold can be aggregated to compute comprehensive evaluation metrics, such as precision, recall, F1-score, ROC AUC, and PR Curve. By maintaining the label distribution, Stratified K-Fold reduces the variance in model evaluation and ensures more reliable and fair comparisons, especially when dealing with imbalanced classification datasets [65].

#### 4. EXPERIMENTS AND RESULT

Table 3 presents the classification performance of five ML models—Decision Tree, XGBoost, Random Forest, LightGBM, and LR—on two prediction tasks: Smart Technologies Readiness and Low-Carbon Industry. Detailed performance metrics are further reported in Table 4. Each model was evaluated using Stratified 5-Fold Cross Validation to ensure that each fold preserved the percentage of samples for each class, thereby producing more balanced and reliable performance estimates across all classes. Evaluation metrics include Accuracy, Macro Precision, Macro Recall, Macro F1-score, PR AUC (Average Precision), and ROC AUC (Average), providing a comprehensive overview of the models' predictive capabilities.

Across both prediction tasks, XGBoost and LightGBM consistently outperformed the other models. XGBoost achieved the highest accuracy in both Smart Technologies Readiness (0.9318) and Low-Carbon Industry (0.9334), accompanied by high Macro F1-scores (0.8512 and 0.7981) and outstanding ROC AUC values (0.9724 and 0.9652). These results underscore the model's robustness in learning complex patterns while maintaining class sensitivity. LightGBM followed closely with similarly high accuracy (0.9306 and 0.9300), as well as high PR AUC and ROC AUC scores, confirming its effectiveness in capturing nonlinear interactions and subtle feature dynamics within industrial readiness data (See Figure 6).

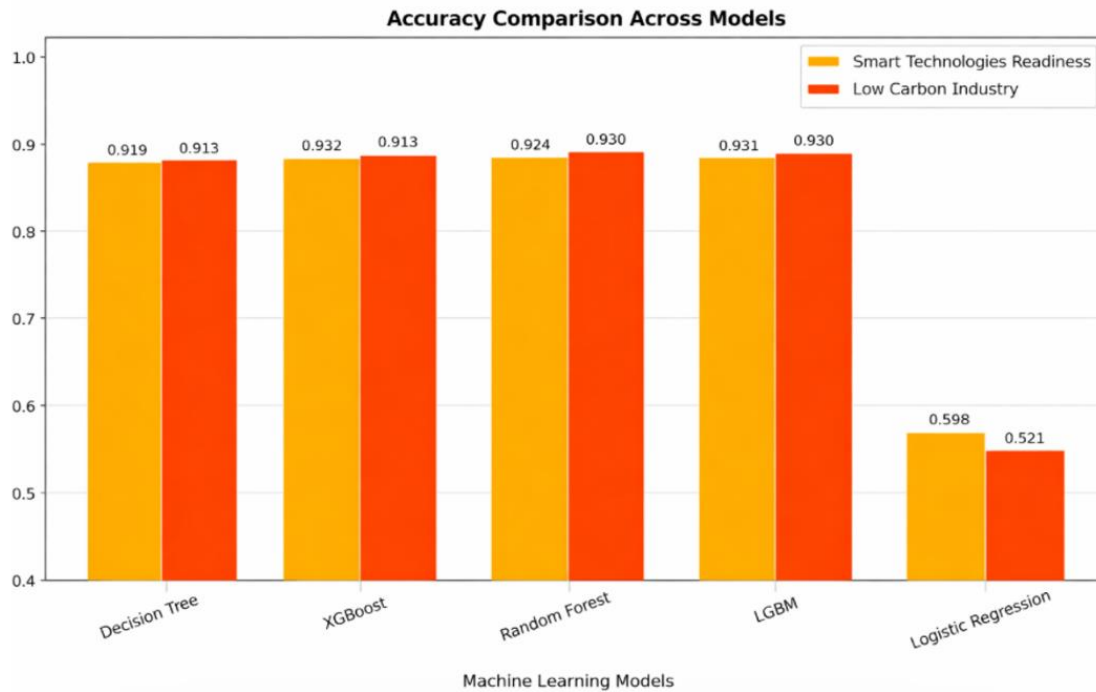
Random Forest demonstrated competitive performance, although slightly lower than the boosting models. It maintained high accuracy (0.9240 and 0.9300) and ROC AUC scores (0.9704 and 0.9626), but a relatively lower recall in the Low-Carbon task (0.6908) indicated some difficulty in capturing the minority class, which could be a concern in imbalanced prediction contexts. Decision Tree performed reasonably well with accuracies of 0.9192 and 0.9126; however, it struggled in terms of precision-recall trade-offs, as seen in its lower PR AUC scores (0.5171 and 0.3489).

These results reflect its limited capacity to model more complex data structures and may explain its weaker performance under stratified sampling. LR consistently showed the lowest performance across all metrics, with accuracy scores of 0.5984 and 0.5206 and PR AUC values below 0.14. These findings suggest that linear models are not suitable for this classification task, likely due to their inability to capture nonlinearity and interaction effects without intensive pre-processing or feature engineering.

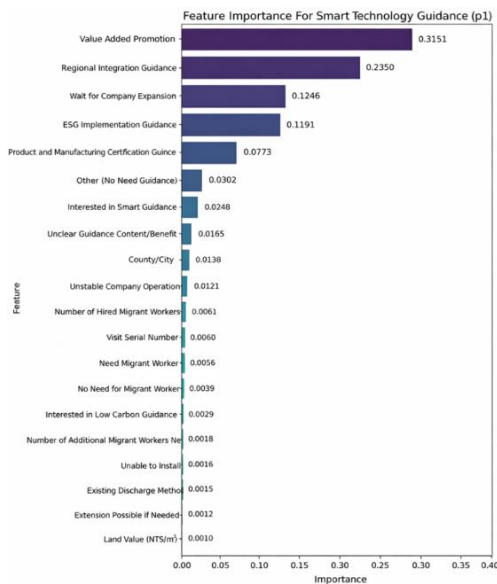
**Table 4.** Classification report on smart technologies readiness and low-carbon industry using machine learning (ML)

| Smart Technologies Readiness |        | Low Carbon Industry        |        |
|------------------------------|--------|----------------------------|--------|
| <b>DT</b>                    |        |                            |        |
| Accuracy                     | 0.9192 | Accuracy                   | 0.9126 |
| Macro Precision              | 0.8207 | Macro Precision            | 0.7522 |
| Macro Recall                 | 0.8238 | Macro Recall               | 0.7507 |
| Macro F1-score               | 0.8219 | Macro F1-score             | 0.7508 |
| PR AUC (Average Precision)   | 0.5171 | PR AUC (Average Precision) | 0.3489 |
| ROC AUC (Average)            | 0.8238 | ROC AUC (Average)          | 0.7507 |
| <b>LGBM</b>                  |        |                            |        |
| Accuracy                     | 0.9306 | Accuracy                   | 0.9300 |
| Macro Precision              | 0.8412 | Macro Precision            | 0.8098 |
| Macro Recall                 | 0.8625 | Macro Recall               | 0.7758 |
| Macro F1-score               | 0.8510 | Macro F1-score             | 0.7901 |
| PR AUC (Average Precision)   | 0.7732 | PR AUC (Average Precision) | 0.7253 |
| ROC AUC (Average)            | 0.9713 | ROC AUC (Average)          | 0.9636 |
| <b>LR</b>                    |        |                            |        |
| Accuracy                     | 0.5984 | Accuracy                   | 0.5206 |
| Macro Precision              | 0.4978 | Macro Precision            | 0.5237 |
| Macro Recall                 | 0.4958 | Macro Recall               | 0.5671 |
| Macro F1-score               | 0.4536 | Macro F1-score             | 0.4300 |
| PR AUC (Average Precision)   | 0.1325 | PR AUC (Average Precision) | 0.1255 |
| ROC AUC (Average)            | 0.4978 | ROC AUC (Average)          | 0.5724 |
| <b>RF</b>                    |        |                            |        |
| Accuracy                     | 0.9240 | Accuracy                   | 0.9300 |
| Macro Precision              | 0.8408 | Macro Precision            | 0.8590 |
| Macro Recall                 | 0.8121 | Macro Recall               | 0.6908 |
| Macro F1-score               | 0.8248 | Macro F1-score             | 0.7419 |
| PR AUC (Average Precision)   | 0.7567 | PR AUC (Average Precision) | 0.7159 |
| ROC AUC (Average)            | 0.9704 | ROC AUC (Average)          | 0.9626 |
| <b>XGB</b>                   |        |                            |        |
| Accuracy                     | 0.9318 | Accuracy                   | 0.9334 |
| Macro Precision              | 0.8468 | Macro Precision            | 0.8208 |
| Macro Recall                 | 0.8566 | Macro Recall               | 0.7814 |
| Macro F1-score               | 0.8512 | Macro F1-score             | 0.7981 |
| PR AUC (Average Precision)   | 0.7751 | PR AUC (Average Precision) | 0.7294 |
| ROC AUC (Average)            | 0.9724 | ROC AUC (Average)          | 0.9652 |

Note: DT = Decision Tree; XGB = Extreme Gradient Boosting (XGBoost); RF = Random Forest; LGBM = Light Gradient Boosting Machine; LR = Logistic Regression; PR AUC = Area Under the Precision-Recall Curve; ROC AUC = Area Under the Receiver Operating Characteristic Curve.



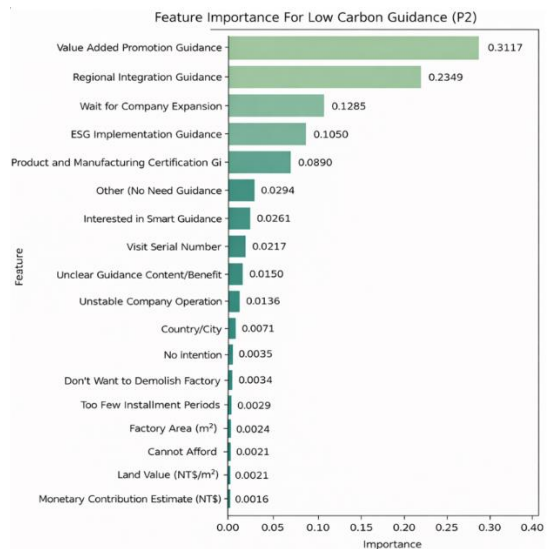
**Figure 6.** Accuracy comparison across models



**Figure 7.** Boruta ranking importance for smart technology guidance

While the results without feature selection already demonstrate the effectiveness of gradient boosting methods in modelling industrial readiness, these models are still trained on the full set of features, some of which may be irrelevant or redundant. The inclusion of such features can potentially introduce noise, increase computational complexity, and reduce the interpretability of the models. To address this, the Boruta feature selection algorithm is employed in the next stage of analysis. By identifying and retaining only the most relevant features, Boruta aims to enhance model performance and robustness. The selected features highlight the importance of technological guidance, renewable energy readiness, and environmental compliance factors. These results indicate that industrial transformation is influenced not only by technological capabilities but also by regulatory and infrastructural conditions. The subsequent section presents the

classification results using the same set of algorithms, now trained exclusively on the features selected by Boruta as shown in Figure 7 and Figure 8, providing a comparative insight into the impact of feature selection on predictive accuracy and generalization.



**Figure 8.** Boruta ranking importance for low carbon guidance

Following the baseline evaluation using the full feature set, feature selection was conducted using the Boruta algorithm to identify and retain only the most relevant predictors. Table 5 summarizes the classification results after applying Boruta for both Smart Technologies Readiness and Low-Carbon Industry targets. Boruta aims to improve model generalization by eliminating redundant or irrelevant features that may introduce noise or increase the risk of overfitting, particularly in datasets with a relatively large number of variables.

Overall, the application of Boruta had a stabilizing effect on model performance across most algorithms. Gradient boosting

models such as XGBoost and LightGBM continued to demonstrate strong predictive capabilities, maintaining high levels of accuracy and macro F1-scores. Their superior performance can be attributed to their ability to capture complex non-linear relationships and feature interactions through sequential tree boosting. By iteratively correcting errors from previous trees, these models can adaptively focus on the most informative patterns within the dataset, which becomes even more effective when irrelevant features are removed.

**Table 5.** Classification report on smart technologies readiness and low-carbon industry using Boruta feature selection

| Smart Technologies Readiness |        | Low Carbon Industry |        |
|------------------------------|--------|---------------------|--------|
| <b>DT</b>                    |        |                     |        |
| Accuracy                     | 0.9204 | Accuracy            | 0.9208 |
| Macro Precision              | 0.8272 | Macro Precision     | 0.7749 |
| Macro Recall                 | 0.8160 | Macro Recall        | 0.7708 |
| Macro F1-score               | 0.8209 | Macro F1-score      | 0.7720 |
| PR AUC                       |        | PR AUC              |        |
| (Average Precision)          | 0.5158 | (Average Precision) | 0.3902 |
| ROC AUC                      |        | ROC AUC             |        |
| (Average)                    | 0.8160 | (Average)           | 0.7708 |
| <b>LGBM</b>                  |        |                     |        |
| Accuracy                     | 0.9252 | Accuracy            | 0.9198 |
| Macro Precision              | 0.8315 | Macro Precision     | 0.7762 |
| Macro Recall                 | 0.8462 | Macro Recall        | 0.7446 |
| Macro F1-score               | 0.8382 | Macro F1-score      | 0.7587 |
| PR AUC                       |        | PR AUC              |        |
| (Average Precision)          | 0.7426 | (Average Precision) | 0.6660 |
| ROC AUC                      |        | ROC AUC             |        |
| (Average)                    | 0.9678 | (Average)           | 0.9563 |
| <b>LR</b>                    |        |                     |        |
| Accuracy                     | 0.9026 | Accuracy            | 0.5216 |
| Macro Precision              | 0.7860 | Macro Precision     | 0.5248 |
| Macro Recall                 | 0.9440 | Macro Recall        | 0.5703 |
| Macro F1-score               | 0.8341 | Macro F1-score      | 0.4312 |
| PR AUC                       |        | PR AUC              |        |
| (Average Precision)          | 0.6603 | (Average Precision) | 0.1256 |
| ROC AUC                      |        | ROC AUC             |        |
| (Average)                    | 0.9583 | (Average)           | 0.5740 |
| <b>RF</b>                    |        |                     |        |
| Accuracy                     | 0.9232 | Accuracy            | 0.9208 |
| Macro Precision              | 0.8240 | Macro Precision     | 0.7804 |
| Macro Recall                 | 0.8529 | Macro Recall        | 0.7361 |
| Macro F1-score               | 0.8373 | Macro F1-score      | 0.7540 |
| PR AUC                       |        | PR AUC              |        |
| (Average Precision)          | 0.7534 | (Average Precision) | 0.6655 |
| ROC AUC                      |        | ROC AUC             |        |
| (Average)                    | 0.9682 | (Average)           | 0.9568 |
| <b>XGB</b>                   |        |                     |        |
| Accuracy                     | 0.9254 | Accuracy            | 0.9224 |
| Macro Precision              | 0.8345 | Macro Precision     | 0.7840 |
| Macro Recall                 | 0.8372 | Macro Recall        | 0.7607 |
| Macro F1-score               | 0.8355 | Macro F1-score      | 0.7703 |
| PR AUC                       |        | PR AUC              |        |
| (Average Precision)          | 0.7354 | (Average Precision) | 0.6779 |
| ROC AUC                      |        | ROC AUC             |        |
| (Average)                    | 0.9676 | (Average)           | 0.9578 |

Note: DT = Decision Tree; XGB = Extreme Gradient Boosting (XGBoost); RF = Random Forest; LGBM = Light Gradient Boosting Machine; LR = Logistic Regression; PR AUC = Area Under the Precision-Recall Curve; ROC AUC = Area Under the Receiver Operating Characteristic Curve.

Random Forest also benefited from the feature selection process, achieving improved macro F1-scores and consistently high ROC AUC values [66]. As an ensemble model based on

bagging, Random Forest reduces variance by aggregating predictions from multiple decision trees trained on different subsets of the data and features. This structure makes it inherently robust to noisy variables and helps maintain stable predictive performance even when the feature space changes [67].

LR showed a notable improvement in the Smart Technology readiness task after feature selection, achieving a substantial increase in accuracy and recall. This improvement suggests that removing irrelevant features helped reduce noise in the linear decision boundary, allowing the model to better capture the dominant patterns in the data. However, its weaker performance in the Low-Carbon task indicates that the underlying relationships in that target variable may be more complex and non-linear, which linear models struggle to represent effectively.

Meanwhile, the Decision Tree classifier exhibited only marginal improvements after feature selection. While decision trees are capable of capturing non-linear relationships, their single-tree structure makes them more susceptible to instability and overfitting compared to ensemble-based methods. This limitation may explain why its overall predictive performance remained lower than that of ensemble models even after irrelevant variables were removed.

Taken together, the Boruta feature selection process effectively refined the input feature space by identifying the most relevant predictors while eliminating less informative variables. This process not only reduced potential noise in the dataset but also enhanced the interpretability of the modelling results. The findings demonstrate that ensemble-based learning algorithms, particularly XGBoost, LightGBM, and Random Forest, are well-suited for industrial readiness prediction tasks that involve heterogeneous and survey-based data. These models can capture complex, non-linear relationships and interactions among multiple predictors, allowing them to achieve strong predictive performance even in the presence of diverse input variables.

Furthermore, the results emphasize the critical role of feature selection in balancing model accuracy and interpretability. By focusing on the most influential variables, the modelling framework provides clearer insights into the key factors associated with industrial readiness for digital and low-carbon transformation. From a policy perspective, the findings suggest that improving industrial readiness requires not only technological adoption but also stronger institutional support and financial mechanisms. Such support systems can facilitate firms' ability to invest in digital infrastructure, adopt low-carbon technologies, and gradually transition toward more sustainable and digitally integrated production systems.

## 5. CONCLUSIONS

This study evaluates the predictive performance of five ML models—Decision Tree, XGBoost, Random Forest, LightGBM, and LR—on two industrial policy targets: Smart Technologies Readiness and Low-Carbon Industry. The models were trained and validated using the full feature set and subsequently re-evaluated using features selected through the Boruta feature selection algorithm. All experiments were conducted using 5-fold Stratified Cross-Validation to ensure balanced representation of each class.

The results demonstrate that ensemble-based models, particularly XGBoost and LightGBM, consistently achieve

superior predictive performance across both tasks in terms of accuracy, precision, recall, and F1-score. The application of the Boruta feature selection method further improved model performance in several cases, especially for Smart Technologies Readiness, indicating that removing irrelevant or redundant features can enhance model generalization and efficiency. In addition, Boruta improves model interpretability by identifying the most relevant variables influencing industrial readiness.

From a policy perspective, these findings highlight the importance of digital capability in supporting the transition toward low-carbon industrial systems. Strengthening digital infrastructure, promoting smart manufacturing initiatives, and encouraging the adoption of digital technologies may help accelerate industrial decarbonization efforts. Furthermore, the ML framework proposed in this study can assist policymakers in identifying critical indicators and prioritizing industrial sectors that require targeted support for digital transformation and sustainable development. For example, variables related to workforce capability and access to technological guidance indicate the importance of knowledge support and training programs. Similarly, variables related to renewable energy infrastructure highlight the need for policies that facilitate access to clean energy resources for industrial firms.

Since the dataset used in this study is cross-sectional, the analysis captures industrial readiness at a single point in time. As a result, the model reflects the current state of digital and low-carbon transformation readiness but does not account for how these conditions evolve over time. Future research could address this limitation by incorporating panel or time-series data to examine the dynamic progression of industrial transformation across sectors. Integrating temporal industrial data and external policy indicators would also enable more robust dynamic prediction and scenario analysis.

In addition, future studies could enhance the interpretability of the predictive framework by applying explainable artificial intelligence techniques such as SHAP and LIME. These approaches would help reveal how individual features contribute to model predictions, thereby improving transparency and making the analytical insights more actionable for policymakers. Such developments would strengthen the practical relevance of ML-based industrial readiness assessments and support evidence-based policy formulation for digital and low-carbon industrial transitions.

## ACKNOWLEDGMENT

The authors would like to express their sincere appreciation to the Industrial Technology Research Institute (ITRI), Taiwan, for providing the dataset used in this study. The authors also acknowledge the support from Honchita Co., Ltd. (Uniform Number: 60304896), Changhua, Taiwan, Chaoyang University of Technology, Satya Wacana Christian University, Universitas Ciputra, Atma Jaya Catholic University of Indonesia, and Bina Nusantara University for their academic and administrative assistance throughout the completion of this research.

## REFERENCES

- [1] Adha, R., Hong, C.Y., Agrawal, S., Li, L.H. (2023). ICT, carbon emissions, climate change, and energy demand nexus: The potential benefit of digitalization in Taiwan. *Energy & Environment*, 34(5): 1619-1638. <https://doi.org/10.1177/0958305X221093458>
- [2] Wu, C.H., Chou, C.W., Chien, C.F., Lin, Y.S. (2024). Digital transformation in manufacturing industries: Effects of firm size, product innovation, and production type. *Technological Forecasting and Social Change*, 207: 123624. <https://doi.org/10.1016/j.techfore.2024.123624>
- [3] Yin, K., Miao, Y., Huang, C. (2024). Environmental regulation, technological innovation, and industrial structure upgrading. *Energy & Environment*, 35(1): 207-227. <https://doi.org/10.1177/0958305X221125645>
- [4] Xu, J., Chen, D., Liu, R., Zhou, M., Kong, Y. (2021). Environmental regulation, technological innovation, and industrial transformation: An empirical study based on city function in China. *Sustainability*, 13(22): 12512. <https://doi.org/10.3390/su132212512>
- [5] Jónsdóttir, S., Gísladóttir, G. (2023). Land use planning, sustainable food production and rural development: A literature analysis. *Geography and Sustainability*, 4(4): 391-403. <https://doi.org/10.1016/j.geosus.2023.09.004>
- [6] Chen, Y., Wang, Z., You, K., Zhu, C., Wang, K., Gan, M., Zhang, J. (2024). Trends, drivers, and land use strategies for facility agricultural land during the agricultural modernization process: Evidence from Huzhou City, China. *Land*, 13(4): 543. <https://doi.org/10.3390/land13040543>
- [7] Ling, X., Gao, Y., Wu, G. (2023). How does intensive land use affect low-carbon transition in China? New evidence from the spatial econometric analysis. *Land*, 12(8): 1578. <https://doi.org/10.3390/land12081578>
- [8] Xinfu, T., Jinglin, L. (2022). Study of the mechanism of digitalization boosting urban low-carbon transformation. *Frontiers in Environmental Science*, 10: 982864. <https://doi.org/10.3389/fenvs.2022.982864>
- [9] Pan, Y.C., Chiu, Y.H. (2017). Formulating assessment indices and strategies for the transition to local industrial development in Taoyuan City, Taiwan. *Sustainability*, 9(3): 364. <https://doi.org/10.3390/su9030364>
- [10] Lin, S.H., Wang, D., Huang, X., Zhao, X., Hsieh, J.C., Tzeng, G.H., Li, J.H., Chen, J.T. (2021). A multi-attribute decision-making model for improving inefficient industrial parks. *Environment, Development and Sustainability*, 23(1): 887-921. <https://doi.org/10.1007/s10668-020-00613-4>
- [11] Ma, Z., Fan, X., Zhang, Y., Hu, B. (2023). Understanding the influencing factors of enterprise transformation and upgrading capability: A case study of the national innovation demonstration zones, China. *Sustainability*, 15(3): 2711. <https://doi.org/10.3390/su15032711>
- [12] Yin, P.Y., Yen, A.Y., Chao, S.E., Day, R.F., Bhanu, B. (2022). A machine learning-based ensemble framework for forecasting PM2.5 concentrations in Puli, Taiwan. *Applied Sciences*, 12(5): 2484. <https://doi.org/10.3390/app12052484>
- [13] Gu, G., Wu, B., Zhang, W., Lu, R., Feng, X., Liao, W., Pang, C., Lu, S. (2023). Comparing machine learning methods for predicting land development intensity. *PLoS One*, 18(4): e0282476. <https://doi.org/10.1371/journal.pone.0282476>
- [14] Rocha, C.F., Quandt, C., Deschamps, F., Cruzara, G. (2025). Digital transformation readiness in large manufacturing firms: A building block model proposition. *Journal of Manufacturing Technology*

- Management, 36(1): 45-68. <https://doi.org/10.1108/JMTM-12-2023-0544>
- [15] Chirumalla, K., Oghazi, P., Nnewuku, R.E., Tuncay, H., Yahyapour, N. (2025). Critical factors affecting digital transformation in manufacturing companies. *International Entrepreneurship and Management Journal*, 21(1): 54. <https://doi.org/10.1007/s11365-024-01056-3>
- [16] Heffron, R., Körner, M.F., Wagner, J., Weibelzahl, M., Fridgen, G. (2020). Industrial demand-side flexibility: A key element of a just energy transition and industrial development. *Applied Energy*, 269: 115026. <https://doi.org/10.1016/j.apenergy.2020.115026>
- [17] Lyu, Y., Liu, Y., Guo, Y., Sang, J., Tian, J., Chen, L. (2022). Review of green development of Chinese industrial parks. *Energy Strategy Reviews*, 42: 100867. <https://doi.org/10.1016/j.esr.2022.100867>
- [18] Shahpari, S., Eversole, R. (2024). Planning to 'hear the farmer's voice': An agent-based modelling approach to agricultural land use planning. *Applied Spatial Analysis and Policy*, 17(1): 115-138. <https://doi.org/10.1007/s12061-023-09538-7>
- [19] Effiong, C., Ngang, E., Ekott, I. (2024). Land use planning and climate change adaptation in river-dependent communities in Nigeria. *Environmental Development*, 49: 100970. <https://doi.org/10.1016/j.envdev.2024.100970>
- [20] Jiang, Y., Yao, G., Xu, J., Tian, Y. (2021). Study in driving strategy and analysis of sustainable and symbiosis development relationship between agricultural industrial clusters and agricultural logistics industry. *Sustainability*, 13(24): 13800. <https://doi.org/10.3390/su132413800>
- [21] Tang, H., Liu, X., Li, M. (2024). Research on the influencing factors of Chinese agricultural brand competitiveness based on DEMATEL-ISM. *Scientific Reports*, 14(1): 11363. <https://doi.org/10.1038/s41598-024-62068-1>
- [22] Hurlimann, A., March, A., Bush, J., Moosavi, S., Browne, G.R., Warren-Myers, G. (2024). Climate change transformation in built environments—A policy instrument framework. *Urban Climate*, 53: 101771. <https://doi.org/10.1016/j.uclim.2023.101771>
- [23] Sakthivel, S., Virmani, C., Priscila, S.S., Pathak, R., Surendhar, S.P.A., Sobirov, B. (2025). Smart technical control infrastructures in electrical automation through digital application systems. *International Journal of Critical Infrastructures*, 21(1): 1-23. <https://doi.org/10.1504/IJCIS.2025.143947>
- [24] Raut, P.K., Das, J.R., Gochhayat, J., Das, K.P. (2022). Influence of workforce agility on crisis management: Role of job characteristics and higher administrative support in public administration. *Materials Today: Proceedings*, 61: 647-652. <https://doi.org/10.1016/j.matpr.2021.08.121>
- [25] Zhou, Q., Wu, J., Imran, M., Nassani, A.A., Binsaeed, R.H., Zaman, K. (2023). Examining the trade-offs in clean energy provision: Focusing on the relationship between technology transfer, renewable energy, industrial growth, and carbon footprint reduction. *Heliyon*, 9(10): e20271. <https://doi.org/10.1016/j.heliyon.2023.e20271>
- [26] Cimino, A., Longo, F., Mirabelli, G., Solina, V., Verteramo, S. (2024). An ontology-based, general-purpose and Industry 4.0-ready architecture for supporting the smart operator (Part II—Virtual Reality case). *Journal of Manufacturing Systems*, 73: 52-64. <https://doi.org/10.1016/j.jmsy.2024.01.001>
- [27] Awouda, A., Traini, E., Bruno, G., Chiabert, P. (2024). IoT-Based Framework for Digital Twins in the industry 5.0 era. *Sensors*, 24(2): 594. <https://doi.org/10.3390/s24020594>
- [28] Vrontis, D., Christofi, M., Pereira, V., Tarba, S., Makrides, A., Trichina, E. (2023). Artificial intelligence, robotics, advanced technologies and human resource management: A systematic review. *The International Journal of Human Resource Management*, 33(6): 1237-1266. <https://doi.org/10.1080/09585192.2020.1871398>
- [29] Zhang, Y., Yao, J., Guan, H. (2018). Intelligent cloud resource management with deep reinforcement learning. *IEEE Cloud Computing*, 4(6): 60-69. <https://doi.org/10.1109/MCC.2018.1081063>
- [30] Salihi, A.A., Ibrahim, H., Baharudin, D.M. (2024). Environmental governance as a driver of green innovation capacity and firm value creation. *Innovation and Green Development*, 3(2): 100110. <https://doi.org/10.1016/j.igd.2023.100110>
- [31] Naeem, M.A., Appiah, M., Karim, S., Yarovaya, L. (2023). What abates environmental efficiency in African economies? Exploring the influence of infrastructure, industrialization, and innovation. *Technological Forecasting and Social Change*, 186: 122172. <https://doi.org/10.1016/j.techfore.2022.122172>
- [32] Mehmood, S., Zaman, K., Khan, S., Ali, Z. (2024). The role of green industrial transformation in mitigating carbon emissions: Exploring the channels of technological innovation and environmental regulation. *Energy and Built Environment*, 5(3): 464-479. <https://doi.org/10.1016/j.enbenv.2023.03.001>
- [33] Naegler, T., Buchgeister, J., Hottenroth, H., Simon, S., Tietze, I., Viere, T., Junne, T. (2022). Life cycle-based environmental impacts of energy system transformation strategies for Germany: Are climate and environmental protection conflicting goals? *Energy Reports*, 8: 4763-4775. <https://doi.org/10.1016/j.egy.2022.03.143>
- [34] Valionienė, E., Župerkienė, E. (2024). Exploring the factors that affect the resilience of port organizational ecosystems through a survey of common uncertainties. *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, 18(1): 185-192. <https://doi.org/10.12716/1001.18.01.19>
- [35] Han, N., Um, J. (2024). Risk management strategy for supply chain sustainability and resilience capability. *Risk Management*, 26(2): 6. <https://doi.org/10.1057/s41283-023-00138-w>
- [36] Theofilatos, K., Likothanassis, S., Karathanasopoulos, A. (2012). Modeling and trading the EUR/USD exchange rate using machine learning techniques. *Engineering, Technology & Applied Science Research*, 2(5): 269-272. <https://doi.org/10.48084/etasr.200>
- [37] Fatima, Z., Tanveer, M.H., Voicu, R.C., Chakravarty, S., Ashfaq, M., Khan, M., Zaib, A., Desa, H. (2025). Explainable AI for IOT devices and robotic communication phishing detection: A machine learning approach using LIME and SHAP. *Engineering, Technology & Applied Science Research*, 15(5): 26478-26486. <https://doi.org/10.48084/etasr.11595>
- [38] Wu, H., Wang, X., Wang, X., Su, W. (2024). Based on machine learning model for prediction of CO<sub>2</sub> adsorption

- of synthetic zeolite in two-step solid waste treatment. *Arabian Journal of Chemistry*, 17(2): 105507. <https://doi.org/10.1016/j.arabjc.2023.105507>
- [39] Huertas-García, Á., Martí-González, C., Maezo, R.G., Rey, A.E. (2024). A comparative study of machine learning algorithms for anomaly detection in industrial environments: Performance and environmental impact. In *Trends in Sustainable Computing and Machine Intelligence*, pp. 373-389. [https://doi.org/10.1007/978-981-99-9436-6\\_26](https://doi.org/10.1007/978-981-99-9436-6_26)
- [40] Cucchi, M., Volpi, L., Ferrari, A.M., García-Muiña, F.E., Settembre-Blundo, D. (2023). Industry 4.0 real-world testing of dynamic organizational life cycle assessment (O-LCA) of a ceramic tile manufacturer. *Environmental Science and Pollution Research*, 30(60): 124546-124565. <https://doi.org/10.1007/s11356-022-20601-7>
- [41] Koilo, V., Danielsen, A.V. (2023). Financial performance-based assessment of companies' competitiveness: Evidence from the norwegian shipbuilding industry. *Investment Management and Financial Innovations*, 20(3): 137-151. [https://doi.org/10.21511/imfi.20\(3\).2023.12](https://doi.org/10.21511/imfi.20(3).2023.12)
- [42] Sarker, I.H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3): 160. <https://doi.org/10.1007/s42979-021-00592-x>
- [43] Joshi, S., Sharma, M. (2024). Intelligent algorithms and methodologies for low-carbon smart manufacturing: Review on past research, recent developments and future research directions. *IET Collaborative Intelligent Manufacturing*, 6(1): e12094. <https://doi.org/10.1049/cim2.12094>
- [44] Hu, Z., Li, S. (2023). Innovation-driven policy and low-carbon technology innovation: Research driven by the impetus of national innovative city pilot policy in China. *Sustainability*, 15(11): 8723. <https://doi.org/10.3390/su15118723>
- [45] Hu, Y. (2023). Has the development of the digital economy reduced the growth rate of carbon emissions? *Frontiers in Business, Economics and Management*, 11(2): 113-119. <https://doi.org/10.54097/fbem.v11i2.12568>
- [46] Nordbeck, R., Seher, W., Grüneis, H., Herrnegger, M., Junger, L. (2023). Conflicting and complementary policy goals as sectoral integration challenge: An analysis of sectoral interplay in flood risk management. *Policy Sciences*, 56(3): 595-612. <https://doi.org/10.1007/s11077-023-09503-8>
- [47] Young, J.D., Evans, A.M., Iniguez, J.M., Thode, A., Meyer, M.D., Hedwall, S.J., McCaffrey, S., Shin, P., Huang, C.H. (2020). Effects of policy change on wildland fire management strategies: Evidence for a paradigm shift in the western US? *International Journal of Wildland Fire*, 29(10): 857-877. <https://doi.org/10.1071/WF19189>
- [48] Sabet, E., Yazdani, N., De Leeuw, S. (2017). Supply chain integration strategies in fast evolving industries. *The international Journal of Logistics Management*, 28(1): 29-46. <https://doi.org/10.1108/IJLM-01-2015-0013>
- [49] Kursu, M.B., Jankowski, A., Rudnicki, W.R. (2010). Boruta—a system for feature selection. *Fundamenta informaticae*, 101(4): 271-285. <https://doi.org/10.3233/FI-2010-288>
- [50] Cheng, Q., Wang, X., Wang, S., Li, Y., Liu, H., Li, Z., Sun, W. (2023). Research on a carbon emission prediction method for oil field transfer stations based on an improved genetic algorithm—The decision tree algorithm. *Processes*, 11(9): 2738. <https://doi.org/10.3390/pr11092738>
- [51] Jena, M., Dehuri, S. (2020). DecisionTree for classification and regression: A state-of-the art review. *Informatica*, 44(4): 405-420. <https://doi.org/10.31449/inf.v44i4.3023>
- [52] Lee, W., Lee, J. (2024). Tree-based modeling for large-scale management in agriculture: explaining organic matter content in soil. *Applied Sciences*, 14(5): 1811. <https://doi.org/10.3390/app14051811>
- [53] Tarwidi, D., Pudjaprasetya, S.R., Adytia, D., Apri, M. (2023). An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach. *MethodsX*, 10: 102119. <https://doi.org/10.1016/j.mex.2023.102119>
- [54] Kwaghtyo, D.K., Eke, C.I. (2023). Smart farming prediction models for precision agriculture: A comprehensive survey. *Artificial Intelligence Review*, 56(6): 5729-5772. <https://doi.org/10.1007/s10462-022-10266-6>
- [55] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1): 5-32. <https://doi.org/10.1023/A:1010933404324>
- [56] Chen, R.C., Dewi, C., Huang, S.W., Caraka, R.E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, 7(1): 52. <https://doi.org/10.1186/s40537-020-00327-4>
- [57] Yang, J., Wang, W., Fu, C., Xu, X., Li, Q. (2024). The impact of smart manufacturing demonstration projects on green innovation of Chinese firms: Based on random forest methods. *Heliyon*, 10(7): e28925. <https://doi.org/10.1016/j.heliyon.2024.e28925>
- [58] Karabayir, I., Butler, L., Goldman, S.M., Kamaleswaran, R., Gunturkun, F., Davis, R.L., Webster Ross, G., Petrovitch, H., Masaki, K., Tanner, C.M., Tsvigoulis, G., Alexandrov, A.V., Chinthala, L.K., Akbilgic, O. (2022). Predicting Parkinson's disease and its pathology via simple clinical variables. *Journal of Parkinson's Disease*, 12(1): 341-351. <https://doi.org/https://doi.org/10.3233/JPD-212876>
- [59] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *NIPS'17: in Proceedings of the 31st International Conference on Neural Information Processing*, California, USA, pp. 3149-3157.
- [60] Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5): 1189-1232. <https://doi.org/10.1214/aos/1013203451>
- [61] Benos, L., Tagarakis, A.C., Dolias, G., Berruto, R., Kateris, D., Bochtis, D. (2021). Machine learning in agriculture: A comprehensive updated review. *Sensors*, 21(11): 3758. <https://doi.org/10.3390/s21113758>
- [62] Bewick, V., Cheek, L., Ball, J. (2005). Statistics review 14: Logistic regression. *Critical Care*, 9(1): 112. <https://doi.org/10.1186/cc3045>
- [63] Lin, T.C., Wang, K.J., Sheng, M.L. (2020). To assess smart manufacturing readiness by maturity model: A case study on Taiwan enterprises. *International Journal of Computer Integrated Manufacturing*, 33(1): 102-115. <https://doi.org/10.1080/0951192X.2019.1699255>

- [64] Osisanwo, F.Y., Akinsola, J.E., Awodele, O., Hinmikaiye, J.O., Olakanmi, O., Akinjobi, J. (2017). Supervised machine learning algorithms: Classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, 48(3): 128-138. <https://doi.org/10.5555/3294996.3295074>
- [65] Fontanari, T., Fróes, T.C., Recamonde-Mendoza, M. (2022). Cross-validation strategies for balanced and imbalanced datasets. In *Brazilian Conference on Intelligent Systems*, pp. 626-640. [https://doi.org/10.1007/978-3-031-21686-2\\_43](https://doi.org/10.1007/978-3-031-21686-2_43)
- [66] Yildiz, E.N., Cengil, E., Yildirim, M., Bingol, H. (2023). Diagnosis of chronic kidney disease based on CNN and LSTM. *Acadlore Transactions on AI and Machine Learning*, 2(2): 66-74. <https://doi.org/10.56578/ataiml020202>
- [67] Erçel, S. Akyol, S. (2025). Impact of data preprocessing techniques on the performance of machine learning models for drought prediction. *Acadlore Transactions on AI and Machine Learning*, 4(1): 14-24. <https://doi.org/10.56578/ataiml040102>