

Labour Demand Prediction using Tree-Based Regression Models in the Agricultural Processing Industry of Thailand

KANOGKAN LEEROJANAPRAPA¹, KOMN BHUNDARAK^{2,*}

¹Department of Statistics,
School of Science, King Mongkut's Institute of Technology Ladkrabang,
Ladkrabang, Bangkok 10520,
THAILAND

²Department of Operations Management,
Thammasat Business School, Thammasat University,
2 Prachan Road, Pranakorn, Bangkok 10200,
THAILAND

**Corresponding Author*

Abstract: - Labour is a critical factor in the manufacturing sector, particularly in agricultural processing industries, which constitute a major industrial base in Thailand. Accurate forecasting of labour demand at the factory level is therefore essential for effective production planning and operational continuity. This study aims to identify an appropriate predictive approach to estimate the number of workers required in individual factories using fundamental production-related factors as predictor variables. Three machine learning models—Decision Tree Regressor, Random Forest Regressor, and Gradient Boosting Regressor—were developed and compared. The dataset was divided into 70% for model training and 30% for testing on unseen data. Based on the results of the 10-fold cross-validation performed on the training dataset, pairwise comparisons using paired *t*-tests showed that the Random Forest Regressor (RFR) achieved significantly lower MAE and MAPE values than both the Decision Tree Regressor (DTR) and the Gradient Boosting Regressor (GBR) ($p < 0.05$). However, for RMSE, although RFR significantly outperformed DTR, the difference between RFR and GBR was not statistically significant ($p > 0.05$). Evaluation on the independent test dataset further demonstrated that RFR achieved the best predictive performance and strong generalisation ability, producing the lowest RMSE, MAE, and MAPE values of 98.0732, 34.2280, and 1.5493, respectively. Furthermore, pairwise comparisons using the Wilcoxon signed-rank test confirmed that the absolute prediction errors of RFR were significantly lower than those of both DTR and GBR at the 95% confidence level ($p < 0.05$). Overall, these findings provide statistical evidence that RFR consistently delivered superior predictive performance across both the cross-validation phase and the independent test dataset. The combination of high predictive accuracy, statistical superiority, and robust generalisation capability indicates that RFR is the most appropriate tree-based regression model for labour demand forecasting in agricultural processing factories.

Key-Words: - Agro-processing sector, Decision tree regressor, Gradient boosting regressor, Labour demand, Random forest regressor, Tree-Based Regression.

Received: April 13, 2026. Revised: June 16, 2026. Accepted: July 19, 2026. Published: October 20, 2026.

1 Introduction

Thailand, long characterised as an emerging industrial economy, is currently undergoing a profound demographic transition toward an ageing society. Rapid industrialisation and economic development in recent decades have been accompanied by declining fertility rates and rising life expectancy, resulting in a growing proportion of older adults in the population [1]. Furthermore, demographic trends indicate that Thailand is

transitioning to an ageing society at a faster pace than many developed countries, including the United Kingdom, the United States, and Singapore [2]. As a result, the proportion of the working-age population has steadily declined, creating significant challenges for labour-intensive industries and exacerbating labour shortages across the country [2], [3]. These shortages have become increasingly severe, disrupting production operations and limiting firms' ability to satisfy market demand [3],

[4]. Therefore, understanding the impact of population ageing on the industrial workforce is essential for formulating sustainable labour policies and strengthening Thailand's long-term economic resilience.

The agro-processing industry is one of Thailand's major industrial sectors and is among the country's largest employers. This industry focuses on transforming agricultural products into processed and semi-processed goods, thereby adding value to agricultural outputs, reducing post-harvest losses, and expanding access to both domestic and international markets. Nevertheless, despite technological advancements and increasing levels of mechanisation, the agro-processing industry remains highly labour-intensive and continues to require a large workforce throughout its production processes. In addition, labour requirements in the agro-processing sector are highly sensitive to seasonal fluctuations in agricultural inputs, variability in production volumes, and the need for rapid processing to preserve product quality. Consequently, firms often rely on a mixed workforce comprising permanent and temporary employees, thereby increasing the complexity of labour planning compared with more capital-intensive industries. These challenges are further exacerbated by population ageing and a contracting labour supply, heightening the risk of workforce shortages that could disrupt production continuity. Therefore, precise labour forecasting and industry-specific workforce planning are essential for sustaining productivity, competitiveness, and long-term industrial development in Thailand's agro-processing sector [2], [5].

Accurate labour demand forecasting supports strategic decisions on workforce restructuring, skills development, and automation investments to alleviate labour constraints and enhance productivity [2]. When planners and decision-makers can anticipate future labour availability with reasonable confidence, they are better able to implement effective operational and strategic plans. In situations where labour shortages are expected to constrain existing production systems, increased investment in machinery and automation may be required to reduce labour dependency. Given the inherent uncertainty surrounding future labour supply, high-accuracy forecasting or predicting is essential.

Recently, [6] summarised the main challenges of workforce forecasting in the construction industry through a systematic review, and some of these challenges are also common to other industries, including a lack of standardised forecasting models.

Many existing models are context-specific and not easily generalizable across different regions or construction contexts, limiting their utility. Next, data availability and quality issues can make accurate forecasting difficult. Lastly, technological and methodological limitations mean that many models still rely on traditional statistical methods and have not fully leveraged advanced data analytics or AI, which could improve accuracy and adaptability.

Recently, a predictive model of labour demand using advanced statistical and machine learning techniques, including deep learning, has been developed. Techniques such as ARIMA, Exponential Smoothing, and LSTM are compared for their forecasting capabilities [2]. Recently, [7] identified key factors influencing the planned number of working hours in the construction workforce using a fuzzy set-based model. In addition, labour demand can be estimated by incorporating multiple influencing factors into multivariate regression models, enabling the analysis of complex relationships and the use of time-series data for forecasting. Similarly, [8] proposed a data-driven machine learning model to forecast delivery positions for workforce planning in shop-floor logistics. However, the study focused primarily on developing an application-specific forecasting model rather than providing a systematic evaluation of alternative tree-based learning paradigms or their statistical performance. Furthermore, at a broader economic level, [9] developed a machine-learning approach to labour-market forecasting by integrating macroeconomic indicators, labour-market statistics, and Google search data. Their findings demonstrated that machine learning techniques effectively captured nonlinear labour market dynamics under rapidly changing economic conditions.

However, multivariate regression is particularly useful for predicting labour demand, particularly when dealing with non-linear relationships in labour data [10]. Recently, [11] developed and compared ten machine learning algorithms, including linear regression models, tree-based ensemble methods, and other predictive techniques, to estimate project-level person-hour requirements for State Transportation Agencies (STAs). The comparative analysis demonstrated that the Random Forest regression model achieved the highest predictive performance among the ten algorithms, highlighting the effectiveness of tree-based ensemble learning for workforce forecasting.

Furthermore, the major concern in applying a regression model to address the nonlinear

relationship identified by [12] was the use of a linear regression model to forecast labour demand in the textile and garment industries. In considering modelling techniques for a nonlinear workforce prediction model, Machine learning with tree-based algorithms was employed, as it outperforms linear regression in addressing the nonlinear relationship. such as using Random Forest and Gradient Boosting for predicting workforce needs, including demand forecasting and resource allocation. [13], predicting the productivity of municipality workers [14], predicting labour shortages [15], and workforce forecasting for healthcare services [16], etc.

The performance of Tree-based models, including DTR, is superior to that of deep learning methods, highlighting their robustness across a range of applications [12]. Decision tree regression methods, particularly Random Forests and Gradient Boosting, are competitive with deep learning approaches such as Long Short-Term Memory (LSTM) networks for retail demand forecasting. Various studies indicate that while deep learning models excel in capturing complex patterns and temporal dependencies, tree-based methods often yield superior accuracy in specific contexts. For example, Tree-based models, such as Extra Trees and Random Forests, have demonstrated better performance on metrics such as Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) than LSTM, especially for certain product categories, such as fresh meat [17]. Decision tree methods are effective at handling structured data and can outperform deep learning models when the dataset is not excessively large or complex [18]. In addition, Decision trees offer greater interpretability and require less computational power, making them more accessible for businesses without extensive data science resources [19].

The present study extends our previous research on labour demand prediction in the Thai textile and garment industry by introducing a substantially different research scope, methodology, and evaluation framework. While [12] employed conventional linear regression models to estimate labour demand, the current work proposes a comprehensive tree-based regression framework to capture complex nonlinear relationships among operational variables in the agricultural processing industry without requiring strict distributional assumptions [20], [21].

The contribution of this study extends beyond comparing existing machine learning algorithms. Specifically, this research:

(i) develops a systematic tree-based forecasting framework for labour demand prediction using DTR, RFR, and GBR for industrial operational data in the agro-processing,

(ii) provides a rigorous comparative evaluation by statistical significance testing to ensure the reliability of the comparison of tree-based regression models.

(iii) offers interpretable feature importance analysis and practical insights to support workforce planning and managerial decision-making.

2 Materials and Methods

This section describes the proposed tree-based regression framework, which integrates representative tree-based learning paradigms, statistical hypothesis testing, and feature importance analysis into a unified methodology for labour demand forecasting.

2.1 Tree-Based Regression Models

Tree-based methods partition the feature space into a set of rectangles and then fit a simple model (e.g., a constant) in each [22]. Rather than arbitrarily selecting three algorithms, this study intentionally compares three representative tree-based learning paradigms, namely a single-tree model (Decision Tree Regressor), a bagging-based ensemble model (Random Forest Regressor), and a boosting-based ensemble model (Gradient Boosting Regressor). This design enables a systematic evaluation of the major theoretical categories of tree-based regression methods under the same forecasting problem. The ML regression models were compared using 3 performance metrics: minimum Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE).

2.1.1 Decision Tree Regression (DTR)

A Regression Tree Regressor is a machine learning model that predicts a continuous numerical value by recursively splitting the data into smaller, more homogeneous groups based on input features. The model iteratively divides the dataset into smaller subsets by selecting features that minimise prediction error. This process creates a hierarchical, tree-like structure in which each internal node corresponds to a decision rule based on a specific feature [23]. Moreover, each branch reflects the result of that decision. Each leaf node contains the predicted value, typically the average of the target values in that subset (Figure 1).

The decision tree regressor can be expressed [23] as:

$$\hat{y}_{DT}(x) = \sum_{m=1}^M c_m * I(x \in R_m) \quad (1)$$

where: R_m denotes the m -th region of the input space defined by the tree structure,
 c_m is the average target value of samples in the region R_m ,
 $I(\cdot)$ is the indicator function.

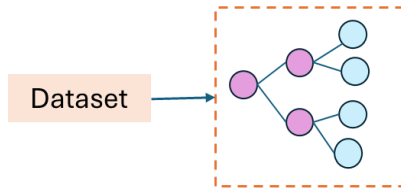


Fig. 1: Decision Tree Regression
Source: created by the authors

2.1.2 Random Forest Regression

Random forests are a substantial modification of bagging that builds a large collection of decorrelated trees, and then averages them [20], [24]. RFR is an ensemble learning model that combines multiple decision trees to predict a continuous target variable. Each tree is trained on a different bootstrap sample of the dataset and makes an independent prediction.

RF uses "bagging" (Bootstrap Aggregating) to build multiple trees in parallel, averaging their results to reduce variance and improve generalizability. Therefore, the random forest predictor [20] can be expressed as: each tree's prediction at the b^{th} tree. Each tree is trained on a bootstrap sample with random feature selection, when implementing RFR (e.g., using RFR in scikit-learn) as shown in Figure 2.

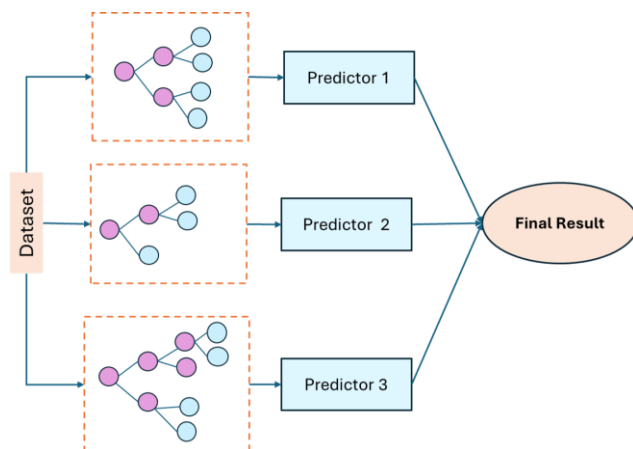


Fig. 2: Random forest regression
Source: created by the authors

RFR employs a method for generating decision trees by randomly sampling both the data and the variables. It then predicts variable values based on the ensemble structure generated from combining the results of each decision tree [25]. This ensemble approach yields highly accurate and stable predictive models. RF effectively handles complex data and intricate relationships between output variables. It splits nodes in each tree using a limited set of features, a process known as bootstrapping [21]. Additionally, it can automatically manage abnormal data or missing values.

A RF is an ensemble learning technique that builds multiple decision-tree regressors on different subsets of the dataset. It then averages their predictions to improve accuracy and minimise overfitting.

$$\hat{f}_{RF}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}_b(x) \quad (2)$$

when:

B is the number of decision trees.

$\hat{f}_b(x)$ is a prediction from the b -th tree.

The trees within the forest use an optimal split strategy, such as setting the splitter to "best" in the "DecisionTreeRegressor". When bootstrap samples are drawn, about one-third of the samples are drawn; the `max_samples` parameter determines the size of each subset; otherwise, each tree is built using the entire dataset (scikit-learn).

2.1.3 Gradient Boosting Regressor

Gradient Boosting Regressor (GBR) is an ensemble machine learning technique that constructs a strong predictive model by sequentially combining multiple weak learners, typically decision trees. The fundamental principle of gradient boosting is to iteratively minimise a predefined loss function by adding new models that correct the residuals of previous models. The model was trained using gradient boosting with decision trees as base learners and optimised through regularised objective minimisation. Shrinkage and subsampling strategies were applied to enhance generalisation and mitigate overfitting. [22] proposed the Gradient Tree Boosting algorithm, which is derived from Friedman's Gradient Boost [26].

As shown in Figure 3, the output values for each leaf of the decision tree are calculated. If a leaf contains multiple residuals, the final output for that leaf is determined by averaging all the values. By iteratively optimising a differentiable loss function via gradient descent in function space, the algorithm primarily reduces model bias and improves

predictive accuracy in the presence of complex nonlinear relationships [26]. This averaging process is applied whether there is a single residual or multiple residuals within the leaf.

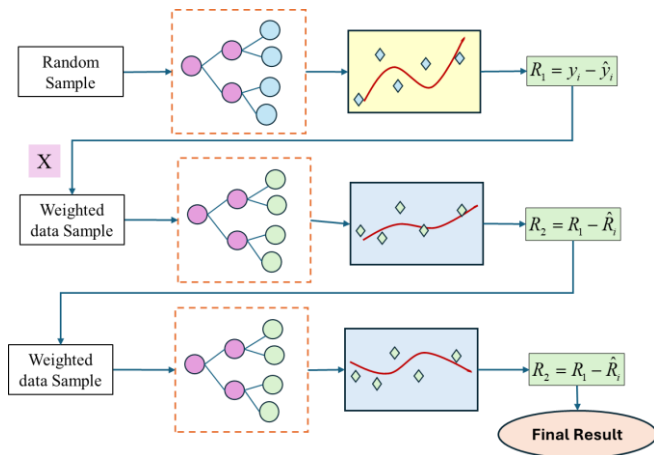


Fig. 3: Gradient boosting regression

Source: created by the authors

Mathematically, this step can be represented as:

$$\gamma_m = \arg \min \sum_{i=1}^n L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) \quad (3)$$

when; L is a loss function,

$h_m(x_i)$ is the DT made on residuals,

$F_{m-1}(x_i)$ is the prediction of the previous model,

γ_i is the predicted value,

m is the number of DT,

M is defined as the last DT.

The leaf's output is the gamma value that minimises the loss function. In this context, the left-hand side "Gamma" represents the output value of a specific leaf.

2.2 Statistical Hypothesis Testing

To compare prediction errors across the machine learning models, hypothesis tests were conducted on the error metrics from each method. The selection of the statistical test was based on the distributional characteristics of the data. First, the normality of the error metrics was assessed using a normality test. In addition, Mauchly's test of sphericity was performed to examine whether the variances of the pairwise differences between models were equal. When the sphericity assumption was violated ($p < 0.05$), the Greenhouse–Geisser correction was applied to adjust the degrees of freedom and control the Type I error rate.

After obtaining a significant overall ANOVA result, parametric statistical tests were employed if

the conditions were satisfied. Conversely, if the normality assumption was violated, non-parametric statistical tests were applied. Accordingly, the statistical methods used in this study are summarised as follows:

2.2.1 Assumptions

The normality of the prediction error metrics was assessed before hypothesis testing. Two statistical tests were considered for evaluating normality: the Shapiro–Wilk (S–W) test and the Anderson–Darling (A–D) test. The Shapiro–Wilk test was applied to datasets with a relatively small sample size, whereas the Anderson–Darling test was employed for larger datasets because of its greater sensitivity to departures from normality, particularly in the tails of the distribution. If the normality assumption was satisfied ($p > 0.05$), parametric statistical methods were used. Otherwise, non-parametric statistical methods were employed. The statistical procedures adopted in this study are summarised below.

Lilliefors' Test

The Lilliefors test is commonly used to assess normality when the population mean and standard deviation are unknown and must be estimated from the sample. While the test applies to a wide range of sample sizes, it is generally recommended for moderate-to-large samples (typically $n > 50$).

$$D = \max |F(X) - S(X)| \quad (4)$$

when; $F(X)$ is the empirical cumulative distribution function (ECDF),

$S(X)$ is the cumulative distribution function of the normal distribution using the sample mean $\hat{\mu}$ and sample standard deviation $\hat{\sigma}$.

Shapiro-Wilk Test

For small to moderate sample sizes (typically $n < 50$), the Shapiro–Wilk test is generally preferred because it demonstrates higher statistical power than the Kolmogorov–Smirnov test in detecting deviations from normality.

$$W = \frac{\sum_{i=1}^n (a_i x_{(i)})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (5)$$

when: n is the number of pairs,
 $x_{(i)}$ is the i^{th} ordered sample value (smallest to largest),
 $\bar{x}$ is the sample mean,

a_i is constants (weights) derived from the expected values and covariance matrix of the order statistics of a standard normal distribution.

2.2.2 Comparison of Repeated Measurements

To compare the predictive performance of the three machine learning algorithms, the number of treatments (or experimental conditions) was employed. Consequently, the prediction error metrics across the three models were not independent but were measured under the same experimental conditions. Repeated Measures ANOVA accounts for this within-fold dependency and provides a valid statistical comparison of the models' predictive performance.

Before conducting the analysis, the assumptions of Repeated Measures ANOVA were evaluated. The normality of each performance metric (MAE, RMSE, and MAPE) across the cross-validation folds was assessed using the Shapiro–Wilk test. A p -value greater than 0.05 indicated that the normality assumption was satisfied.

Repeated Measures ANOVA

The general equation for the F-statistic in a one-way repeated measures ANOVA is:

$$F = \frac{SS_{Treatment} / n - 1}{SS_{Error} / ((k - 1)(n - 1))} \quad (6)$$

When:

$$SS_{Error} = SS_{Total} - SS_{Treatment} - SS_{Subject}$$

$$SS_{Total} = \sum_{j=1}^k \sum_{i=1}^n (Y_{ij} - \bar{Y}_{..})^2$$

$$SS_{Treatment} = n \sum_{j=1}^k (\bar{Y}_{.j} - \bar{Y}_{..})^2$$

$$SS_{Subject} = k \sum_{i=1}^n (\bar{Y}_{i.} - \bar{Y}_{..})^2$$

$$SS_{Treatment} = \frac{\sum_{j=1}^k T_j^2}{n} - \frac{T^2}{kn}$$

when: X_{ij} is the observation i under treatment j ,

T_j is the sum of the observations in treatment (j),

T is the sum of all observations,

$\bar{X}_{..}$ is the grand mean of all observations,

n is the number of samples (or experimental units),

k is the number of treatments (or experimental conditions).

Friedman Test

The Friedman test statistic, as shown by Q in equation (7), is:

$$Q = \left[\frac{12}{nk(k+1)} \sum_{j=1}^k (R_j)^2 \right] - 3n(k+1) \quad (7)$$

when: n is the number of subjects (blocks),

k is the number of treatments,

R_j is the sum of ranks for treatment j .

For moderate or large samples ($n > 10$ or $k > 5$), Q approximately follows a chi-square distribution as $Q \sim \chi_{k-1}^2$.

2.2.3 Pairwise Comparisons

Pairwise comparisons were conducted using paired t -tests with a Bonferroni adjustment to control the family-wise error rate for multiple comparisons. Statistical significance was determined at the 5% level ($p < 0.05$).

Paired t-test

When the two samples are related (dependent), a paired-samples t -test is used to determine whether there is a significant difference between the two population means. The observations should be obtained from a randomly selected sample drawn from a population in which the variable of interest is approximately normally distributed. The dependent variable should be continuous and measured on an interval or ratio scale.

The test statistic of the paired t -test is defined as:

$$t = \frac{\bar{d}}{\frac{S_d}{\sqrt{n}}} \quad (8)$$

when: $\bar{d}$ is the sample mean of the differences,

n is the number of pairs,

S_d is the sample standard deviation of the differences.

The Wilcoxon Signed-Rank Test

The Wilcoxon Signed-Rank Test is a nonparametric alternative to the paired-samples t -test. It is used to determine whether there is a significant difference between two related (dependent) samples when the assumption of normality is violated.

$$W = \text{Minimum} \left| \sum R_i^+, \sum R_i^- \right| \quad (9)$$

For large sample sizes ($n \geq 25$), the test statistics can be approximated by the normal distribution using:

$$Z = \frac{W - \mu_w}{\sigma_w} = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \quad (10)$$

where

n is the number of nonzero paired differences,

R_i is the rank of the absolute difference.

2.3 Feature Importance Analysis

For tree-based regression models, feature importance is commonly computed using the Mean Decrease in Impurity (MDI), also referred to as impurity-based feature importance. During tree construction, each split is selected to maximise the reduction in node impurity. The importance of a predictor is then calculated by summing the impurity reduction attributable to that predictor across all split nodes and, for ensemble methods, averaging these contributions across all trees [20]. The feature importance of the predictor X_j can be expressed as:

$$FI_j = \sum_t \sum_{n \in N_j^{(t)}} \Delta I_n \quad (11)$$

when:

T is the number of trees in the ensemble;

$N_j^{(t)}$ denotes the set of nodes in the tree t that split on the predictor X_j ;

ΔI_n is the impurity reduction achieved by the node n .

The impurity reduction for regression trees is typically computed using the reduction in Mean Squared Error (MSE):

$$\Delta I = I(P) - \frac{N_L}{N_p} I(L) - \frac{N_R}{N_p} I(R) \quad (12)$$

when: $I(P)$ is the impurity of the parent node,

$I(L)$ and $I(R)$ are the impurities of the left and right child nodes,

N_p , N_L and N_R denote the number of observations in the parent, left child, and right child nodes, respectively.

2.4 Methods

All analyses were conducted in Python 3.12. Data preprocessing was performed using Pandas and NumPy. Machine learning models were implemented using Scikit-learn, while the XGBoost algorithm was developed using the XGBoost library. Statistical hypothesis testing was performed using the SciPy library. Model performance was evaluated

using Scikit-learn metrics, and graphical visualisations were produced with Matplotlib. This section presents the methodology used to develop the proposed model.

2.4.1 Data Collection

The manufacturer of Processed Agriculture (ISIC CODE 10xxx) is registered under the Thailand Standard Industrial Classification (TSIC), which the Department of Employment, Ministry of Labour, Thailand, uses to classify industries.

Table 1. Relevant variables in this study

ID	Variable	Description	Unit
1	HP	Horsepower	hp
2	CapLand	Land capital	1,000 Baht
3	CapBuid	Buildings capital	1,000 Baht
4	CapMach	Machinery capital	1,000 Baht
5	Capwork	Working capital	1,000 Baht
6	BArea	Building area	m ²
7	Farea	Factory area	m ²
8	Nui_hp1	Operating horsepower	hp
9	Total	The labour demand of a manufacturer	persons

Source: created by the authors

Table 1 presents the size of manufacturers—measured by factory and building area—as well as their operating capacity, indicated by horsepower, are connected to the number of workers involved in production and the primary horsepower utilised. This study focuses exclusively on Type 3 factories, which are subject to specific regulatory requirements concerning their scale of operation, potential environmental and safety impacts, and the requirement to obtain an operational license prior to commencing activities.

2.4.2 Data Preparation

The data preparation process began with data screening to assess missing values and potential outliers and to verify data accuracy and consistency. Since this study focused on tree-based regression models, no variable transformation or feature scaling was performed to standardise measurement units, unlike in some other machine learning approaches. The dataset used in this study comprised 6,403 industrial factories registered in the industrial database, each with complete records for all variables and no missing values or abnormal observations.

2.4.3 Feature Selection

The variable selection process began by examining the relationships among variables using the Pearson correlation coefficient (r) to assess the strength and direction of linear associations.

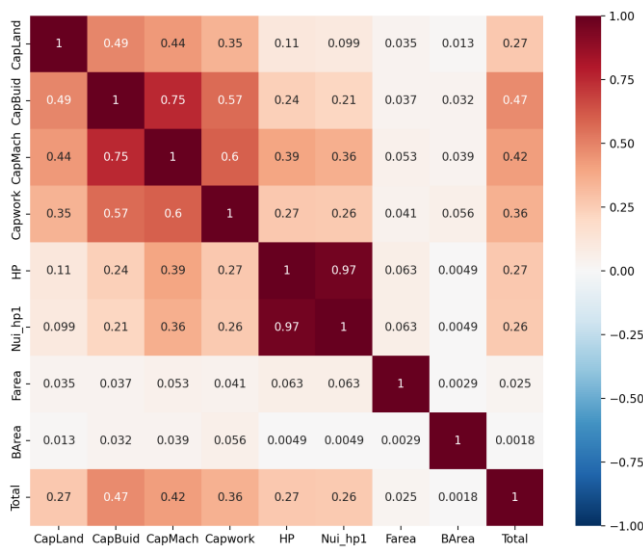


Fig. 4: Feature correlation matrix

Source: created by the authors

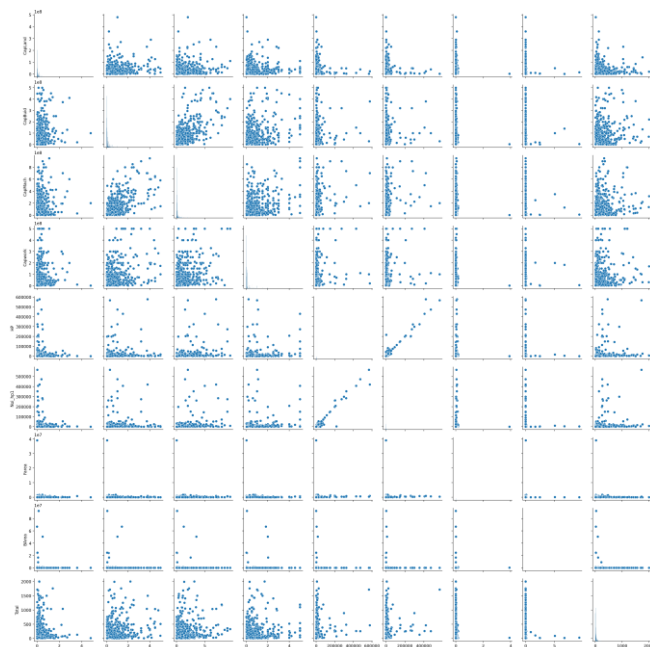


Fig. 5: Relationship data to a higher-dimensional feature space

Source: created by the authors

Figure 4 presents the correlation matrix based on the Pearson correlation coefficient (r) for all pairs of variables, including eight independent variables and one dependent variable (Total). A strong positive correlation was observed between HP and Nui_hp1 ($r = 0.97$), suggesting potential multicollinearity. Although, HP exhibits a slightly stronger correlation with the dependent variable ($r = 0.28$) compared to Nui_hp1 ($r = 0.27$), HP also shows stronger correlations with other independent variables. This suggests that retaining HP may increase the risk of multicollinearity in the model.

Although tree-based regression models are generally robust to multicollinearity, highly correlated variables describing similar operational characteristics may introduce unnecessary redundancy without substantially improving predictive performance. In this study, the Horsepower (HP) variable exhibited a strong correlation with Operating Horsepower (Nui_hp1). Both variables represent machine power characteristics; however, Operating Horsepower (Nui_hp1) more directly reflects the machinery's actual operating conditions during the production process, whereas HP represents the machine's nominal installed capacity. Because labour demand is more closely associated with actual production activities than installed capacity, Operating Horsepower (Nui_hp1) was retained as the more informative predictor. At the same time, HP was excluded to reduce feature redundancy and improve model interpretability. Consequently, the final model was developed using the retained predictor variables, including Operating Horsepower (Nui_hp1). Therefore, HP was excluded from this study's analysis.

Figure 5 indicates nonlinear relationships within the dataset, so the classical linear regression model is limited. Therefore, this provided the rationale for selecting tree-based machine learning models.

2.4.4 Data Split and Data Preprocessing

The dataset was first divided into training and testing sets. Model training and hyperparameter tuning were performed using k-fold cross-validation on the training data, with the test set reserved for final performance evaluation. Randomised Search CV was applied to 50 different hyperparameter sets, with $k = 10$ folds and the training data, where models were trained on nine folds and validated on the remaining fold. The reported performance metrics represent the average values across all folds.

2.4.5 Model Fitting

Model tuning within the defined hyperparameter range was used to determine optimal hyperparameters for each algorithm, and the reported performance metrics represent the average across all folds of the optimal model.

2.4.6 Model Evaluation

The generalisation ability of the proposed models can be evaluated by testing subsets. Model training and hyperparameter tuning were conducted exclusively on the training data using k-fold cross-validation. The trained models were then applied to the unseen testing

data, which were not used at any stage of model development, to obtain an unbiased assessment of predictive performance. The low RMSE and MAE values on the unseen test set indicate that the model maintains strong predictive accuracy on new data.

3 Results

This section presents the results of the tree-based regression models for predicting labour demand. It begins with a descriptive statistical analysis of the dataset, followed by model development on the training dataset and a comparative performance evaluation on the testing dataset. Finally, model interpretability is examined through feature importance analysis.

3.1 Descriptive Statistics

As a preliminary step, descriptive statistics were computed to examine the data before model development. This analysis aimed to summarise the fundamental characteristics of both independent and dependent variables. The data set initially consisted of 10,487 factories; however, only 6,403 had complete records and were therefore retained for analysis. Since all variables were quantitative, the results were presented as mean, standard deviation, minimum, and maximum values, as shown in Table 2.

Table 2. Descriptive statistics for particular features in this study

Variable	Mean	Std.	Min	Max
CapLand	7,498,888.00	17,711,340.00	1,000	480,000,000.00
CapBuid	14,150,070.00	36,937,100.00	1,000	500,000,000.00
CapMach	21,938,700.00	65,592,440.00	2	950,000,000.00
Capwork	12,975,310.00	39,612,490.00	4	500,000,000.00
Nui_hp1	2,317.30	15,995.55	0	566,291.40
Farea	39,312.80	512,120.60	6	39,085,540.00
BArea	51,291.12	1,634,759.00	10	92,411,030.00
Total	43.10	119.27	2	2,000.00

Source: created by the authors

The results in Table 2 indicate substantial variability among factories, particularly in cost-related variables. Notably, machinery cost (A) ranged from 2,000 to 9,500,000,000 baht, with a standard deviation of 65,592,440.00 baht and an average of 21,938,700.00 baht.

3.2 Results from Training Data

The selection of the hyperparameter ranges was guided by the theoretical foundations of regression trees [23] and standard practices in statistical learning [22]. The tuning strategy followed commonly adopted machine learning workflows [24] and the scikit-learn implementation framework.

The hyperparameter ranges for the algorithms are shown in Table 3.

Table 3. Experimental setup of hyperparameter values for 3 models in the experiments.

Model	Variable	Hyperparameters	Optimal
DTR	max_depth	1-20	3
	min_samples_split	2-10	9
	min_samples_leaf	2-10	2
	max_features	Sqrt, log2, None	log2
RFR	n_estimators	10-300	90
	max_depth	10	10
	max_features	Sqrt, log2	log2
	min_sample_leaf	1-6	6
	min_samples_split	2-20	9
GBR	Learning rate	0.01-0.3	0.01
	max_depth	2-10	7
	n_estimators	10-300	140
	subsample	0.2-1.0	0.35
	max_features	Sqrt, log2, None	Sqrt
	min_samples_split	2-10	2
	min_samples_leaf	1-20	2

Source: created by the authors

The performance of DTR, RFR, and GBR models was compared using 10-fold cross-validation. The training dataset was divided into 10 folds: 9 for training and 1 for evaluation. This was repeated 10 times. The model's results were evaluated at each round using RMSE, MAE, and MAPE. After that, the performance criteria for the 10 rounds were averaged for validation across all folds for the evaluated models, as shown in Table 4. RFR achieved the best predictive performance, yielding the lowest RMSE of 101.25, MAE of 0.96, and MAPE of 1.65. The GBR shows moderate performance, while the DT algorithm produced the highest prediction error among all models.

Not only predictive accuracy, but also model stability was evaluated using the standard deviation of performance metrics across cross-validation folds. The results in Table 4 show that the RFR model exhibited relatively low RMSE variability. In contrast, GBR showed slightly lower standard deviations of MAP and MAPE than RFR. Therefore, RFR indicates robust, consistent predictive capability. Based on the cross-validation results, the RFR was selected as the best-performing algorithm. As a result, the optimised RFR model used $n_estimators = 90$, a maximum depth of 10, $min_sample_leaf = 6$, and $min_samples_split = 9$, resulting in improved predictive performance, as shown in Table 3. By treating cross-validation and test datasets as unseen data, RFR achieves better performance, with the lowest MAE, MAPE, and RMSE.

Table 4. MAE, RMSE, and MAPE of 10-fold cross-validation and F-statistic of a repeated measures ANOVA

Feature	Fold	MAE	RMSE	MAPE
DTR	1	37.70	88.63	2.42
	2	42.32	121.91	1.85
	3	29.91	50.90	2.47
	4	50.98	162.46	2.46
	5	38.68	96.14	2.61
	6	41.12	91.46	2.69
	7	49.79	167.26	1.75
	8	42.73	106.57	3.49
	9	38.51	114.18	2.32
	10	35.78	86.30	2.58
	Mean	40.75	108.58	2.46
	Std	5.95	33.51	0.45
RFR	1	31.86	78.77	1.50
	2	40.58	118.67	1.52
	3	26.49	48.99	1.61
	4	45.91	156.99	1.48
	5	35.09	95.27	1.79
	6	36.32	79.37	1.59
	7	46.78	158.76	1.33
	8	39.69	86.10	2.23
	9	34.10	105.51	1.58
	10	34.12	84.07	1.84
	Mean	37.09	101.25	1.65
	Std	5.94	33.18	0.24
GBR	0	34.90	80.83	2.13
	1	42.03	118.72	2.04
	2	28.32	45.06	2.34
	3	47.81	160.81	2.07
	4	36.64	97.97	2.39
	5	37.15	77.35	2.32
	6	49.61	165.43	1.83
	7	37.56	76.10	2.65
	8	35.08	106.39	2.13
	9	35.75	86.08	2.40
	Mean	38.49	101.48	2.23
	Std	6.04	36.08	0.22
F (p-value)		29.2205 (<0.0001)	7.1435 (0.0052)	62.6061 (<0.0001)

Source: created by the authors

Table 5. Paired *t*-tests for pairwise comparison

Compare	MAE	RMSE	MAPE
DTR RFR	8.3775 (<0.0001)	3.8751 (0.0038)	9.0828 (<0.0001)
DTR GBR	4.0901 (0.0027)	2.3872 (0.0407)	2.6221 (0.0277)
RFR GBR	-3.0951 (0.0128)	-0.1537 (0.8813)	-19.0068 (<0.0001)

Source: created by the authors

The predictive performance of the three machine learning models was compared using repeated measurements obtained from the 10-fold cross-validation procedure. Before hypothesis testing, the normality assumption was assessed using the Shapiro–Wilk test. The results indicated that the performance metric for each model was

approximately normally distributed (*p*-value > 0.05). Therefore, a repeated-measures ANOVA was conducted.

Table 4 summarised the repeated-measures ANOVA results, which revealed statistically significant differences among the three machine learning models across all 3 performance metrics: MAE, RMSE, and MAPE. Specifically, significant model effects were observed for MAE ($F = 29.2205$, p -value < 0.0001), RMSE ($F = 7.1435$, p -value < 0.0001), and MAPE ($F = 62.6061$, p -value < 0.0001). These results indicate that at least one model differed significantly from the others for each performance metric across the *k*-fold cross-validation procedure.

To identify the specific differences between models, pairwise comparisons were performed using paired *t*-tests with appropriate adjustment for multiple comparisons as shown in Table 5. For MAPE, the Random Forest Regression (RFR) model achieved significantly lower prediction errors than both DTR (p -value < 0.0001) and GBR (p -value < 0.0001). Similarly, for MAE, the RFR model produced significantly lower errors than DTR (p -value < 0.0001) and GBR (p -value = 0.0128). For RMSE, RFR significantly outperformed DTR; however, the difference between RFR and GBR was not statistically significant (p -value = 0.8813). Overall, these findings demonstrate that the RFR model consistently achieved the best predictive performance across the *k*-fold cross-validation procedure.

3.3 Results from Model Comparison

The predictive accuracy of the proposed models was further evaluated using the test dataset to examine their generalisation capability under unseen conditions.

Table 6 presents the comparative performance of the developed models on the test and unseen datasets. The results also show that the RFR outperformed the DT and GBR models, achieving the lowest RMSE, MAE, and MAPE values of 98.0732, 34.2280, and 1.5493, respectively. This indicates that the RF model achieves higher prediction accuracy and stronger generalisation for predicting labour demand.

Table 6. Performance Comparison based on RMSE, MAE, and MAPE from the test dataset

Model	RMSE	MAE	MAPE
DR	105.6371	39.0873	2.3303
RFR	98.0732	34.2280	1.5493
GBR	99.2150	35.7063	2.1058

Source: created by the authors

Although the cross-validation results demonstrated strong predictive performance, evaluation on the test dataset provides a more realistic assessment of model generalisation. The test results are consistent with the cross-validation performance, indicating that the model does not suffer from significant overfitting.

Before comparing the predictive performance of the 3 models, the distribution of the absolute errors on the test set was examined using the Anderson–Darling normality test. The results indicated that the absolute errors of all three models significantly deviated from a normal distribution (p-value < 0.001). Therefore, the assumption of normality was violated, rendering parametric repeated measures tests inappropriate. Consequently, the non-parametric Friedman test was employed to compare the overall performance of the three models, followed by pairwise Wilcoxon signed-rank tests with Holm adjustment.

Table 7. Friedman test for the absolute prediction errors

Model	Mean	Median	Q Statistics
DTR	39.0873	18.3329	372.1281 (<0.0001)
RFR	34.2280	11.0733	
GBR	35.7063	16.1802	

Source: created by the authors

Table 8. Pairwise Wilcoxon signed-rank test for the absolute prediction errors

Absolute error	W Statistics	p-value
DTR RFR	609,870	<0.0001
DTR GBR	755,525	<0.0001
RFR GBR	587,224	<0.0001

Source: created by the authors

The Friedman test, Table 7, revealed a statistically significant difference among the three machine learning models ($\chi^2 = 372.1281$, p-value < 0.0001), indicating that at least one model exhibited significantly different prediction errors. The mean and median of the absolute error of RF were the lowest, at 34.2280 and 11.0733, respectively. Since lower absolute error indicates better predictive performance, RFR achieved the lowest mean rank and demonstrated the best overall performance.

Pairwise comparisons were subsequently performed using the Wilcoxon signed-rank test, as Table 8. The results showed that RFR significantly outperformed DT (p-value < 0.0001) and GBR (p-value < 0.0001). Furthermore, GBR also achieved significantly lower prediction errors than DTR (p-value < 0.0001). Since lower absolute error

indicates better predictive performance, the RFR achieved the lowest mean rank. The subsequent Wilcoxon signed-rank tests further confirmed that RFR significantly outperformed both the DTR and GBR in terms of absolute prediction error at the 95% confidence level.

3.4 Key Predictive Features

The analysis results for identifying the important variables contributing to each model's predictions were obtained using feature importance derived from the training dataset.

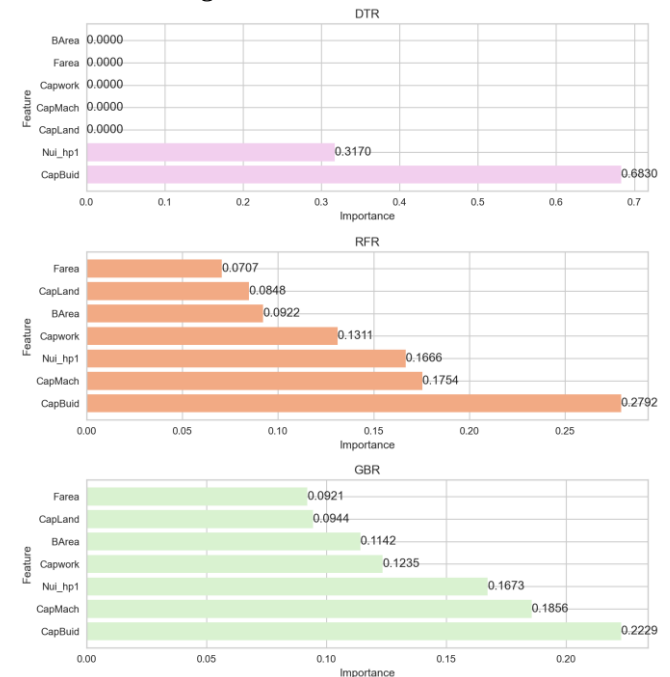


Fig. 6: Feature importance across three algorithms.
Source: created by the authors

Figure 6 presents the feature importance derived from the three algorithms. Overall, Buildings capital (CapBuild) consistently ranked as the most influential predictor across all algorithms, indicating its strong contribution to the target variable. In contrast, the ranking of feature importance in RFR and GBR is the same, while the range of feature importance values varies from the lowest of Factory area (Farea) to the highest of Buildings capital (CapBuild), showing a wider range in FTR, 0.0707 - 0.2792, than in GBR, 0.0921-0.2229. Meanwhile, the feature importance analysis indicates that only two features, namely Buildings capital (CapBuild) and Operating horsepower (Nui_hp1), contribute substantially to the DT model's predictive performance. The remaining variables exhibit negligible importance, suggesting a limited impact on the target variable.

4 Conclusion

This study demonstrates the application of a theoretically grounded and statistically rigorous tree-based regression framework for labour demand prediction, combining representative tree-learning paradigms, hypothesis testing for model comparison, and interpretable feature importance analysis to support industrial workforce planning. Three algorithms—Decision Tree Regressor (DTR), Random Forest Regressor (RFR), and Gradient Boosting Regressor (GBR)—were compared to determine the most suitable model. Previous studies have indicated that DTR can outperform traditional statistical methods, such as multivariate regression, particularly when modelling nonlinear relationships in labour data [10]. Similarly, tree-based models, including DTR, have shown superior performance compared to deep learning approaches in retail demand forecasting, highlighting their robustness across various applications [17]. However, despite its advantages, DTR is prone to overfitting and requires careful parameter tuning to achieve optimal performance. The results of this study revealed that the RFR provided the best predictive performance on both the training and testing datasets. This superior performance can be attributed to its bagging technique, which builds multiple decision trees independently from random subsets of data and averages their outputs to obtain the final prediction. In contrast, the GBR employs a boosting approach that builds trees sequentially, with each new model attempting to correct the errors of the previous one. Consequently, RFR is generally less susceptible to overfitting noisy data than GBR, which may sometimes fit noise while attempting to minimise residual errors.

RFR consistently outperformed the DTR and GBR in terms of predictive accuracy and stability. This superior performance can be attributed to its bagging-based ensemble strategy, which combines bootstrap sampling with random feature selection to reduce prediction variance and improve generalisation performance [20]. In contrast, although Gradient Boosting effectively reduces model bias by sequentially fitting residuals [27], its iterative learning process may increase sensitivity to noisy observations and overfitting, particularly when applied to moderately sized industrial datasets. Since labour demand is influenced by complex operational factors and unavoidable measurement uncertainty, the robustness of RFR enables it to capture nonlinear relationships while maintaining stable predictive performance across different validation folds. RF addresses this limitation by employing bootstrap aggregating (bagging), in

which multiple decision trees are independently trained on different bootstrap samples of the training data. The final prediction is obtained by averaging the predictions of all trees, thereby reducing the variance associated with individual trees and improving model stability and generalisation performance [20], [27]. In addition, a notable limitation of the dataset obtained from factory registration systems is the potential inaccuracy of the reported information. The recorded data may not fully reflect actual operational conditions, as firms may have underlying incentives related to government benefits, financial support, or tax obligations that influence how they report information. Consequently, such discrepancies may introduce systematic patterns or conditional biases in the data. In this context, rule-based tree models are particularly suitable, as they can capture nonlinear relationships and structured patterns in the data, making them well-suited to predicting labour demand at the firm level.

Acknowledgement:

The authors would like to express their sincere gratitude to the Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, for providing research funding under Research Grant No. [2563-02-05-11].

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work, the author(s) utilized artificial intelligence (AI) technologies solely for the purpose of linguistic refinement and grammatical correction. Following the application of these tools, the author(s) comprehensively reviewed, verified, and edited the resulting text. The author(s) maintain full accountability and responsibility for the accuracy, integrity, and final content of the publication.

References:

- [1] United Nations, Department of Economic and Social Affairs, "Population Division, World Population Prospects 2022: Summary of Results." 2022. New York, NY, USA: United Nations, 2022, [Online]. https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/wpp2022_summary_of_results.pdf (Accessed Date: January 15, 2026).
- [2] A. Singh, "Thailand's aging population hits consumption and economic growth," Thailand Business News, Jun. 3, 2024, [Online].

- https://www.thailand-business-news.com/economics/142662-thailands-aging-population-hits-consumption-and-economic-growth?utm_source=chatgpt.com. (Accessed Date: January 15, 2026).
- [3] World Bank. "Aging and the labor market in Thailand," June 22, 2021, [Online]. <https://www.worldbank.org/en/country/thailand/publication/aging-and-the-labor-market-in-thailand> (Accessed Date: January 15, 2026).
- [4] World Bank. "Thailand economic monitor: Aging society and economy" June 2016. Washington, D.C.. World Bank Group, [Online]. <https://documents.worldbank.org/en/publications/documents-reports/documentdetail/830261469638312246> (Accessed Date: January 15, 2026).
- [5] ASEAN Briefing, "The rise of Southeast Asia as a global mobility hub," Oct. 23, 2025, [Online]. Available: <https://www.aseanbriefing.com/news/the-rise-of-southeast-asia-as-a-global-mobility-hub/>. (Accessed Date: January 15, 2026).
- [6] Y. Zhao, K. Qi, A. P. Chan, Y. H. Chiang, and M. F. F. Siu, "Manpower forecasting models in the construction industry: A systematic review," *Engineering, Construction and Architectural Management*, Vol. 29, No. 8, 2022, pp. 3137–3156, <https://doi.org/10.1108/ECAM-05-2020-0351>.
- [7] E. Plebankiewicz and P. Karcińska. "Model for supporting construction workforce planning based on the theory of fuzzy sets," *Applied Sciences*, Vol. 14, No. 4, 2024, pp. 1655, <https://doi.org/10.3390/app14041655>.
- [8] P. Eichenseer, L. Hans, and H. Winkler, "A data-driven machine learning model for forecasting delivery positions in logistics for workforce planning," *Supply Chain Analytics*, Vol. 9, 2025, pp. 100099, <https://doi.org/10.1016/j.sca.2024.100099>.
- [9] J. M. Orozco-Castañeda, L. P. Sierra-Suárez, and P. Vidal, "Labor market forecasting in unprecedented times: A machine learning approach," *Bulletin of Economic Research*, Vol. 76, No. 4, 2024, pp. 893–915, <https://doi.org/10.1111/boer.12451>.
- [10] S. Mühlbauer and E. Weber, "Machine learning for labor market matching," *Machine Learning with Applications*, 2026, pp.100861, <https://doi.org/10.1016/j.mlwa.2026.100861>.
- [11] A. Ogungbire and S. K. Mitra, "Workforce forecasting for state transportation agencies: A machine learning approach," *International Journal of Transportation Science and Technology*, Vol. 17, 2025, pp. 345–360, <https://doi.org/10.1016/j.ijtst.2024.05.004>.
- [12] K. Leerojanaprapa and K. Bhundarak. "Resource factors affecting labour demand for textile and garment industry in Thailand," *Journal of Industrial Engineering and Management Science*, Vol. 1, 2017, pp. 263-288.
- [13] A. K. Kalusivalingam, A. Sharma, N. Patel, & V. Singh, "Optimising workforce planning with AI: leveraging machine learning algorithms and predictive analytics for enhanced Decision-Making," *International Journal of AI and ML*, Vol. 1 No. 3, 2020, pp. 1-27.
- [14] R. M. Sousa, A. P. da Silva, M. A. Alves, and J. P. de Souza, "Predicting the productivity of municipality workers: a comparison of six machine learning algorithms," *Economies*, Vol. 12, No. 1, Art. no. 16, 2024, <https://doi.org/10.3390/economies12010016>.
- [15] A. Williams, "Predicting labor shortages from labor demand and labor supply data: A machine learning approach," *arXiv preprint arXiv:2004.01311*, 2020, <https://arxiv.org/abs/2004.01311>.
- [16] D. T. Koç and E. Eren, "Workforce forecasting with machine learning for healthcare management," *Journal of Health Management*, Vol. 28, No. 1, 2026, pp. 121–135, <https://doi.org/10.1177/09720634251396339>.
- [17] M. Nasser, T. Falatouri, P. Brandtner, and F. Darbanian, "Applying machine learning in retail demand prediction—A comparison of tree-based ensembles and long short-term memory-based deep learning," *Applied Sciences*, Vol. 13, No. 19, 2023, pp. 11112, <https://doi.org/10.3390/app13191112>.
- [18] A. ul Husna, S. H. Amin, and A. Ghasemipoor, "Demand forecasting using machine learning and deep learning approaches in the retail industry: A comparative study," in *Calisir, F., Durucu, M. (eds) Industrial Engineering in the Covid-19 Era. GJCIE 2022. Lecture Notes in Management and Industrial Engineering.*, Cham, Switzerland: Springer, 2023, pp. 260–265, https://doi.org/10.1007/978-3-031-25847-3_24.
- [19] M. S. Shak, A. Uddin, M. H. Rahman, N. Anjum, M. N. V. Al Bony, M. Alam, M. Helal, A. Khan, P. Das, and T. Pervin, "Innovative machine learning approaches to foster financial inclusion in microfinance,"

International Interdisciplinary Business Economics Advancement Journal, Vol. 5, No. 11, 2024, pp. 6–20, <https://doi.org/10.55640/business/volume05issue11-02>.

- [20] L. Breiman, "Random forests," *Machine learning*, Vol. 45, No. 1, 2001, pp. 5–32, <https://doi.org/10.1023/A:1010933404324>.
- [21] C. Hegde and K. E. Gray, "Use of machine learning and data analytics to increase drilling efficiency for nearby wells," *Journal of Natural Gas Science and Engineering*, Vol. 40, 2017, pp. 327–335, <https://doi.org/10.1016/j.jngse.2017.02.019>.
- [22] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009, ISBN: 978-0-387-84858-7.
- [23] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. CRC press, 1984. ISBN: 9781315139470.
- [24] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed. O'Reilly Media, Inc, 2022. ISBN: 9781098122478.
- [25] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," in *Ensemble Machine Learning: Methods and Applications*, New York: Springer, 2012, pp. 157–175. <https://doi.org/10.1007/978-1-4419-9326-7>.
- [26] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of statistics*, Vol. 29, No. 5, 2001, pp. 1189–1232.
- [27] L. Breiman, "Bagging predictors," *Machine Learning*, Vol. 24, No. 2, 1996, pp. 123–140, <https://doi.org/10.1007/BF00058655>.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

- Kanogkan Leerojanaprapa: carried out Model construction with python, investigation, Methodology, Software, Data visualisation and Statistical testing.
- Kom Bhundarak: Investigation, Methodology, Data curation, Statistical testing, Conceptualization and Review & editing.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

This research was funded by Faculty of Science, King Mongkut's Institute of Technology Ladkrabang under Grant No. [2563-02-05-11].

Conflict of Interest The authors have no conflicts of interest to declare.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0 https://creativecommons.org/licenses/by/4.0/deed.en_US