MJFAS
Malaysian Journal of Fundamental and Applied Sciences

Shabli *et al.* | Malaysian Journal of Fundamental and Applied Sciences, Vol. 21 (2025) 2771–2789

RESEARCH ARTICLE

Improving the Prediction of Labour Skill Classification Model in Malaysia using Tree-based Machine Learning Algorithms

Rabi'atul'adawiah Shabli[a,b], Ahmad Zia Ul-Saufie[a]*, Nurain Ibrahim[a]

[a]School of Mathematical Sciences, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA 40450 Shah Alam, Selangor, Malaysia; [b]Department of Statistics Malaysia, Ministry of Economy, Federal Government Administrative Centre, Putrajaya 62514, Wilayah Persekutuan Putrajaya, Malaysia

Abstract Understanding the skill level among the workforce is essential for analysing the quality of the labour market and supporting economic development. The objective of this study is to classify skill level of the Malaysian workforce into skilled, semi-skilled and low-skilled categories using supervised machine learning techniques. Current approaches rely on descriptive statistics which limit the capability of interaction between multiple features and the prediction of future outcomes. As the labour market scenario has shifted towards the adoption of digitalisation and automation, it is essential to adopt more effective and robust methods to identify key factors influencing skill level. This study applied the Cross-Industry Standard Process for Data Mining (CRISP-DM) process to analyse 120,518 cases from the 2023 Salaries and Wages Survey dataset. The dataset undergoes a comprehensive data preprocessing procedure of data cleaning, data transformation, data splitting and handling multiclass imbalanced data by leveraging the Synthetic Minority Oversampling Technique (SMOTE). Five tree-based algorithms were applied including Decision Tree, Random Forest, Gradient Boosted Trees, Adaptive Boosting and Extreme Gradient Boosting which is consistently recognised for their strong classification performance. Model performance was evaluated using four metrics including specificity, sensitivity, F1-score and accuracy. Random Forest achieved the best performance with an accuracy of 86.45%, sensitivity of 86.45%, specificity of 90.89% and F1-score of 86.36%. The findings indicated that Random Forest is effective in predicting the skill level category. Relevant factors contributing to the prediction were salaries and wages received, economic activity, education level, certificate obtained and year of birth. It provides valuable insights into enhancing skill development initiatives that contribute to academic research by applying machine learning techniques in labour market studies.

Keywords: Classification, machine learning, skill level, SMOTE, tree-based algorithm.

***For correspondence:**
ahmadzia101@uitm.edu.my

Received: 15 May 2025
Accepted: 2 Sept. 2025

Introduction

Understanding the skill level within the workforce is important to develop strong human capital, particularly for developing countries such as Malaysia. The Malaysian labour market continues to face challenges relating to underemployment where an individual is working below their qualifications, mismatch between skill and job requirements and reliance on low-skilled [1]. As the current labour market landscape has shifted towards automation and digitalisation, it is necessary to create a workforce equipped with advanced skills and able to adapt working in a fast-changing technological environment [1].

To address this, it is important to determine the scope of labour and understand the categorisation of skill level. Labour refers to any kind of work carried out by an individual using physical and mental contributions to produce goods and services [2]. According to economic theory, labour is part of the

factors of production, together with land, capital and entrepreneurship [3]. As of 2023, there were 15.8 million workers in the Malaysian labour market with 9.7 million male and 6.1 million female [4]. Occupational classifications are grouped into skill level where the development of labour skills is important to enhance productivity growth and economic development [5]. In Malaysia, the occupational classification follows the Malaysia Standard Classification of Occupations (MASCO) 2020, which is aligned with the International Standard Classification of Occupations (ISCO-08) under the International Labour Organization [6].

MASCO 2020 categorised occupations into ten major groups which are further categorised into three levels of skill including skilled, semi-skilled and low-skilled labour [6] using statistical approaches [4]. The current approach utilises descriptive statistics that solely depend on a single categorical feature which is occupation [4]. However, descriptive statistics do have several limitations including its limited scope, inability to establish a causal relationship, unable for analytical analysis, tendency towards biased data and over-reliance on the summary measures which is a constraint on its efficiency for evidence-based policymaking [7].

To overcome this limitation, machine learning is an efficient technique as it allows for exploratory analysis without relying on any empirical model [8] as it learns from previous data to identify patterns in large, unstructured and complex databases [9]. Among several machine learning techniques, classification approaches are suitable for predicting target categories from input features [10]. In recent years, the application of tree-based algorithms has grown as it is capable of handling interactions between multiple input features and the target variable with the ability to manage the non-linear relationship [11] which makes this technique suitable to classify skill level.

From a global perspective, machine learning approaches have been widely employed in labour market classification studies. However, in the Malaysian context, the application of machine learning particularly in categorising labour skills and identifying factors influencing skill level at the national level by using household survey data has not yet been implemented. Existing studies in other fields such as healthcare [12], [13], [14] and sports [15], [16] have successfully utilised supervised machine learning approaches like Decision Tree, Random Forest, Support Vector Machine, k-Nearest Neighbours and Neural Network to handle classification issues.

Another challenge in labour skill classification is the issue of an imbalanced dataset where the distribution of the target variable namely the skill level is not equally represented [17]. In an imbalanced dataset, the model tends to focus and become biased toward the majority classes which is the semi-skilled category while it performed poorly for the underrepresented classes [18]. In this dataset, the semi-skilled category dominated the class distribution whereas skilled and low-skilled are underrepresented in the dataset [19]. Hence, the oversampling technique namely Synthetic Minority Oversampling Technique (SMOTE) is a method to address imbalanced classes by constructing synthetic data for minority classes [20]. Additionally, the existing approach does not incorporate feature selection methods to reduce the dimension of the large input predictors and enhance model accuracy [21].

As of now, there are limited studies predicting skill level as existing literature concentrates on skill level in medical, sport and garment activities. In the Malaysian context, studies on labour market analysis focus on graduate employability, job classification and occupational injuries. Therefore, this study aims to develop a classification model to predict Malaysian skill level into skilled, semi-skilled and low-skilled categories based on features included in the household survey data. Five tree-based algorithms will be employed to assess the model's performance and identify relevant features including Decision Tree, Random Forest, Gradient Boosted Trees, Adaptive Boosting and Extreme Gradient Boosting. Decision Tree is recognised for its simplicity and interpretability, the Random Forest for its robustness in handling imbalanced data, Gradient Boosted Trees for capturing complex patterns, Adaptive Boosting for its efficiency on weak learners and Extreme Gradient Boosting for its accuracy and fast execution. The selection was based on the suitability for managing both categorical and continuous features with its strong capability and interpretability in single trees, ensemble and boosting techniques. Moreover, the operation of tree-based algorithms does not require assumptions on the underlying dataset with the ability to capture non-linear relationships between input predictors and the target variable which contrasts with traditional Logistic Regression [22]. These algorithms have proven to be an efficient method for classification tasks in various domains and a powerful technique for predicting skill level for further exploration. The outcomes of this study offer valuable insights into strengthening labour market modelling and support the initiatives focusing on the importance of skills among the workforce at the national level.

Related Works

Several studies have been carried out to assess labour market information from the perspective of statistical and machine learning approaches. In terms of statistical approaches, Logistic Regression is widely used in managing binary classification tasks. This method was employed to determine the structures of labour market participation in Italy focusing on the second-generation youth [23], predicting factors influencing the overeducated individuals in Turkey [24], to define occupational skill requirements in Germany [25] and evaluate the relationship among the status of employment and the mental health outcomes [26]. Additionally, [27] utilised Logistic Regression to select factors impacting the vertical and horizontal mismatches among young graduates in Hungary whereas [28] applied it to determine the employment likelihood among construction workers in Turkey. In conclusion, Logistic Regression is a widely common statistical approach to solve binary classification that provides reliable information towards identifying important factors and understanding the association between the target variable and input variables [29]. However, [29] highlighted its limitations on non-linear relationships and are prone to overfitting.

Machine learning models are excellent in determining relationships and complex data structures, particularly in dealing with a large number of features [30]. There were various statistical and machine learning approaches which have been applied in the labour market classification studies. A literature by [31] classified graduates' employability among Technical and Vocational Education institutions in Malaysia by employing Logistic Regression, Neural Network and Random Forest with Random Forest achieving the highest accuracy. Furthermore, [32] employed machine learning to assist graduates in selecting appropriate sectors and companies based on their location and interests with five models employed including Decision Tree, Random Forest, Naïve Bayes, Support Vector Machines and k-Nearest Neighbours. Random Forest outperformed other models with 99.4% accuracy. In the meantime, [33] applied Logistic Regression, Random Forest and Extreme Gradient Boosting to explore factors influencing fatal occupational injuries among migrant workers with Extreme Gradient Boosting achieving the highest accuracy. These machine learning approaches have been effectively utilised in various disciplines focusing on prediction accuracy [34], [35].

The existing statistical approach in classifying the skill level within the Malaysian context is by utilising descriptive statistics with considering a single feature which is the major occupational group according to MASCO 2020 [4]. It does not take into account any other factors influencing the skill level classification. There are limitations as it does not capture any interaction between multiple features, has limited scope and cannot predict the future skill level [7]. Thus, existing approaches may not be suitable to meet the current labour market scenario as rapid technological changes will drive economic growth. As of now, the Malaysian labour market continues to face challenges of skill mismatch and underutilisation of skilled labour. Rapid changes in technological advancement and industry requirements have made the labour market more dynamic. Currently, there are no extensive applications of machine learning approaches emphasising skill level at the national level despite its capability of handling large datasets, enhancing classification accuracy, identifying the structure of datasets, predicting skill shortages and the ability to automate statistical analysis.

Tree-based algorithms are widely used in solving classification issues as it is capable of handling non-linear relationships and interactions between target features and input features [11]. In addition, it can manage mixed types of data comprising categorical and continuous data, can handle large predictors and is easy to understand. It concentrates on identifying the most important variables to improve the prediction [36]. Tree-based algorithms consist of single-tree and ensemble approaches namely the bagging and boosting techniques.

The Decision Tree under the single tree is a non-parametric method that is suitable for predicting quantitative or qualitative information [37]. Random Forest, a bagging ensemble approach is built on the foundation of the Decision Tree [38] able to handle large datasets, missing values and outlier data [36]. Meanwhile, the Gradient Boosted Trees offers high accuracy in the model prediction and executes tasks rapidly [39] while Adaptive Boosting uses a Bayesian classifier to minimise the error in the prediction model by integrating several weak models to achieve a powerful model [40]. Another powerful boosting technique is Extreme Gradient Boosting with the ability to mitigate overfitting and obtain high accuracy [33].

Overall, the tree-based algorithm is useful for classification issues in which this technique needs to balance between accuracy, interpretability and computational efficiency. This study fills this gap by applying five tree-based algorithms to determine the most effective algorithm to classify skill level and identify the most relevant features contributing to the classification of skill level.

Methods

The methodology to predict the skill level classification is based on a quantitative approach through employing supervised machine learning approaches. This study adopts the Cross Industry Standard Process for Data Mining (CRISP-DM) as it is accessible to the public and provides a standardised structure which is suitable for data mining in both academic research and industrial applications [41]. The research framework is illustrated in Figure 1. This study utilised the Salaries and Wages Survey dataset which undergoes data preprocessing procedures and continued with the classification process using tree-based algorithms. Next, measuring model verification using performance metrics namely specificity, sensitivity, F1-score and accuracy. Finally, an optimal model was chosen as the proposed model for skill level classification according to the best evaluation of the model's performance.

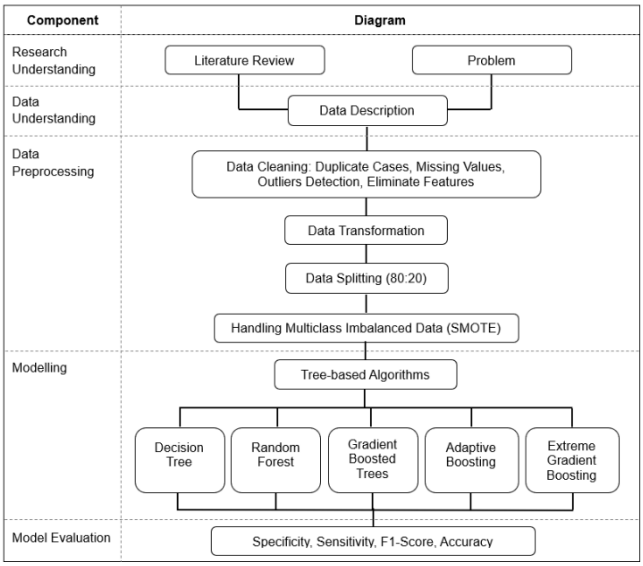

Figure 1. Research Framework using CRISP-DM

Data Description

This study utilises a secondary dataset namely the Salaries and Wages Survey for the reference year 2023, a household survey approach collected by the Department of Statistics Malaysia [19]. It follows the international guideline, An Integrated System of Wages Statistics [42] published by the International Labour Office and adopts a two-stage stratified sampling design [19]. The dataset consists of 120,518 cases with a combination of 53 features in both categorical (47 features) and continuous (6 features) data types. The categorical features refer to economic activity, certificate obtained, state of birth, education level, employment status, gender and marital status whereas continuous features comprise salaries & wages received, working hours and age. The target variable is the skill level which can be classified into three categories namely skilled, semi-skilled and low-skilled labour.

Data Preprocessing

The process of data preprocessing involves four procedures as depicted in Figure 2. It starts with data cleaning, followed by data transformation, data splitting and handling multiclass imbalanced data. These procedures are to ensure data quality and readiness for model prediction to assure a robust analysis [43].

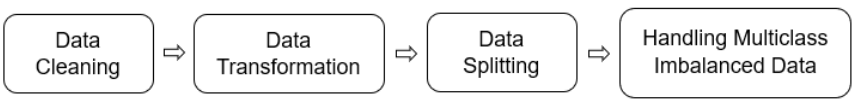

Figure 2. Process of Data Preprocessing

Data Cleaning

First, the data cleaning procedure consists of managing duplicate cases, missing values on features, eliminating irrelevant features, missing values for cases and outlier detection. There were five duplicate cases with similar identification numbers and the cases have been removed from the dataset [44]. It was observed that there were incomplete data with 12 features recorded with missing values. The missing values were examined descriptively by calculating the percentage of incomplete entries across individual features. Nine features involving 7 categorical features and 2 continuous features with more than 30 per cent of missing values were removed to ensure data quality and consistency [45] as it contribute noise to the dataset [32]. Additionally, seven features were removed as it contains similar components with different levels of reporting to prevent duplicate and irrelevant information.

Data imputation is conducted for two cases with less than 20 per cent of missing values from the complete data [46] involving other benefits received from employers and working hours. Mode imputation was employed to the categorical feature namely other benefits received from employers whereas median imputation was applied to the continuous feature of working hours due to its skewness value of 0.62. Data trimming on outlier values for two continuous features was performed by excluding cases within the 1st and 99th percentiles to prevent the impact of extreme values [47]. Upon completion of the data cleaning procedure, 116,747 cases with 38 features remained for further analysis.

Data Transformation

Data transformation is a process of converting selected features into an appropriate format for the purpose of statistical analysis [48]. This study employed ordinal encoding techniques to transform six features including strata, ethnic, citizenship, marital status, education level and certificate obtained into structured criteria suitable for analysis.

Data Partitioning

Data partitioning in this study used the data splitting technique with a ratio of 80:20. The dataset consists of 116,747 observations and has been partitioned randomly into 80% of training dataset and the remaining 20% of validation dataset [9], [37] as illustrated in Figure 3. This procedure enables the evaluation of the model generalisation on the unseen dataset and mitigating overfitting or underfitting issues [49]. The training dataset allows the algorithm to learn and recognise patterns whereas the validation dataset is to evaluate the algorithm performance and update the weights to improve accuracy [50].

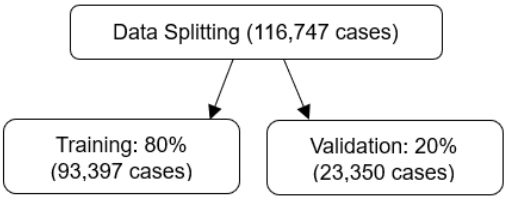

Figure 3. Data Partitioning using the Splitting Method

Handling Multiclass Imbalanced Data

To address the imbalanced classification issue within the target variable while enhancing the model accuracy, the Synthetic Minority Oversampling Technique (SMOTE) was implemented only to the training dataset comprising 93,397 cases as depicted in Figure 4. SMOTE reduces classification bias and helps prevent overfitting to the majority classes [52] by generating synthetic data on the minority classes. The equation of SMOTE is defined by [52] using Equation 1.

$$f_{new} = f_i + (f_i - f_{near})\ X\ R \tag{1}$$

where f_{new} is a newly generated sample, f_i is a randomly selected sample from the minority class, f_{near} is a randomly selected nearest neighbour of f_i from the minority class and R is a random number ranging from 0 and 1 to examine the degree of the interpolation.

In this study, SMOTE was employed solely on the training dataset to obtain a balanced distribution across the skill level classes. The model evaluation was conducted on the original imbalanced testing

dataset to avoid bias in synthetic data. This confirms SMOTE efficiently addressed imbalanced classes during training without disturbing the reliability of the model evaluation outcomes [53].

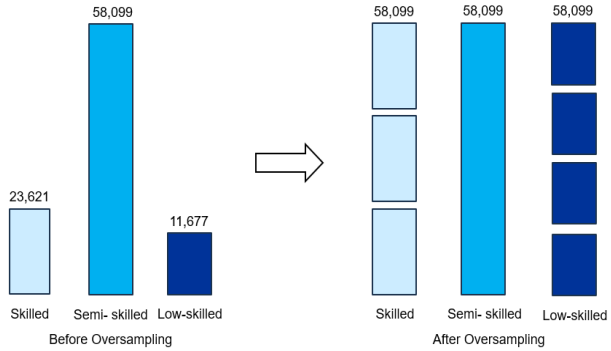

Figure 4. Diagram of SMOTE on Training Dataset

Tree-Based Classification Algorithms

Tree-based algorithms employ a hierarchical structure by splitting data into branches to generate predictions [54]. These algorithms are typically known for their ability to handle tabular data which is widely employed in real-world applications [54]. This study employed five tree-based algorithms to identify features influencing skill level within the Malaysian workforce. The selected algorithms include Decision Tree, Random Forest, Gradient Boosted Trees, Adaptive Boosting and Extreme Gradient Boosting. The outcome of these algorithms will be compared to determine the best-proposed model according to performance measurement for predicting skill level classification [37].

Decision Tree

A Decision Tree is a single-tree algorithm which constructs predictions using a tree-like structure of nodes, branches and leaf nodes to address regression and classification issues [55]. It is effective in managing categorical and continuous features with the ability to handle incomplete data [39]. The algorithm involves two steps in developing the trees by constructing and pruning the trees.

In this study, `DecisionTreeClassifier` was employed from the `sklearn.tree` module. A random state was fixed at 42 to allow for reliability. The Gini index was used as the default splitting metric to achieve the best split. The algorithm typically chooses splits with lower impurity as it represents a purer node [56]. Measurement of node impurity using the Gini index which ranges between 0 to 1 is defined in Equation 2.

$$Gini = 1 - \sum_{i=1}^{n}(P_i)^2 \quad (2)$$

where n is the total number of unique classes in the dataset and P_i is the proportion of cases that belong to class i. The minimum sample size needed for splitting an internal node was set to 2 (`min_samples_split=2`). Meanwhile, the minimum sample size required for a leaf was set to 1 (`min_samples_leaf=1`). Feature selection was based on the best split while no predefined limit on the tree depth. Pruning and regularisation techniques were disabled during the training process.

Random Forest

Random Forest operates as an ensemble machine learning approach that combines several decision trees using the bagging technique [55]. It is capable of handling missing values, outliers and complex data structures while reducing overfitting [58]. For classification tasks, this algorithm constructs multiple decision trees to predict the outcome classes and the final class is chosen through majority voting among the trees [59].

This study utilised `RandomForestClassifier` from the `sklearn.ensemble` module. The model was set to employ 100 trees with a random state fixed at 42 to ensure consistency. Other parameters implemented were the Gini index as the splitting criterion and automatic feature selection according to the square root of the number of input features (`max_features='sqrt'`). There was no limitation on

the tree depth unless further splitting was constrained by an insufficient number of samples or class purity. Measurement of Gini impurity is using Equation 2 [56]. Meanwhile, the Random Forest can be measured based on Equation 3 [60].

$$RF = argmax_{j\epsilon\{1,2,\dots C\}} \sum_{i=1}^{i} DT_{i,j} \tag{3}$$

where $argmax_{j\epsilon\{1,2,\dots C\}}$ is the majority voting function, i is the number of decision trees and j is the number of possible classes in the target features.

Gradient Boosted Trees

Gradient Boosted Trees is a boosting technique categorised under ensemble machine learning algorithms [61]. It uses a loss function to minimise the prediction errors gradually until it achieves the optimal outcome [55]. It is effective in handling non-linear data, noisy data and complex feature interactions [39]. The model constructs the trees sequentially to strengthen model performance by utilising a negative gradient to the loss function through mitigating errors from the previous iterations [62].

The algorithm utilised the `GradientBoostingClassifier` from `sklearn.ensemble` module. The model has been set to perform 100 boosting iterations while the random state was fixed at 42 to ensure reproducibility. A learning rate of 0.1 with a maximum tree depth of 3 was employed for each weak learner. This algorithm implemented `log_loss` function that is suitable for multiclass classification task. The entire dataset was used for every boosting iteration with all input features included during the splitting procedure. The model was trained without early stopping and manual hyperparameter optimisation. It can be customised by the selection of loss function, weak learner depth and additive model [63] as formulated in Equations 4 and 5.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{4}$$

$$\gamma_m = arg\ min_\gamma \sum_{i=n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) \tag{5}$$

where F_m represents the model at-m from gradient boosting approaches, x is the input value, γ is the coefficient value and h_m is the decision tree at-m.

Adaptive Boosting

Adaptive Boosting operates as an ensemble algorithm by combining several weak models and emphasising correcting the errors from the previous models [18]. It generates robust predictions by sequentially minimising error from the preceding models by combining several weak learners [40].

This study utilised `AdaBoostClassifier` from the `sklearn.ensemble` module. The algorithm constructed 100 boosting iterations with a fixed random state of 42 was employed for statistical consistency. The learning rate was configured to 1.0. The base estimator was implemented using `DecisionTreeClassifier` configured as a decision stump by setting the maximum tree depth to 1. No application of manual hyperparameter tuning was employed during the model development. Adaptive Boosting can be computed using Equation 6 [64].

$$H(x) = sign(\sum_{t=1}^{T} a_t H_t(x)) \tag{6}$$

where $H(x)$ is the outcomes of the developed ensemble model, $a_{1,\dots}$, a_t is a set of weights, T is the total number of iterations or weak learners in the ensemble and $H_t(x)$ is the performance of weak learners $t \in (1, \dots, T)$.

Extreme Gradient Boosting

Extreme Gradient Boosting is a boosting technique under the ensemble approach that is widely recognised for its computational efficiency, capability to achieve higher accuracy and ability to mitigate overfitting [33]. It gradually generates an ensemble of decision trees where every subsequent learner emphasises correcting errors made from preceding trees to increase overall performance [62].

In this study, the Extreme Gradient Boosting algorithm was implemented using `XGBClassifier` from the `xgboost` library. It involves 100 boosting iterations with a learning rate of 0.3. The maximum tree

depth was configured to 6 to minimise overfitting. The objective function was set to `multi:softmax` with `num_class=3` that is suitable for multiclass classification. The evaluation metric used was `mlogloss` while label encoding was disabled. Other parameters including a fixed random state of 42 with both `subsample` and `colsample_bytree` were set to 1.0. There is no manual hyperparameter tuning employed during this process.

The objective function of Extreme Gradient Boosting comprises of loss function for minimising classification error and a regularisation term for considering several leaves and its weights [62]. The measurement can be computed using Equation 7 [62].

$$Objective\ function = \sum_{i=1}^{N}\sum_{k=1}^{k}\mathcal{L}(y_{ik}, P_{ik}) + \sum_{m=1}^{n}\Omega(f_m) \quad (7)$$

where N is the number of data points, k is the number of classes, $\mathcal{L}(Y_{ik}, P_{ik})$ is the cross-entropy loss function for the i-th data point and k-th class, P_{ik} is the predicted probability for a i-th data point to k-th class and $\Omega(f_m)$ is the regularisation term for the m-th weak learner.

Model Evaluation

Model evaluation is an important procedure to understand the skill level classification performance of the model prediction. The classification in this study is segmented into three classes namely Skilled (Class 1), Semi-skilled (Class 2) and Low-skilled (Class 3). A literature by [65] stated that the multiclass classification performance can be measured using a confusion matrix as depicted in Table 1.

Table 1. Confusion Matrix for 3-Class Problems

		Predicted Class		
		Skilled = 1	**Semi-skilled = 2**	**Low-skilled =3**
Actual Class	Skilled = 1	TP	FN	FN
	Semi-skilled = 2	FP	TN	FN
	Low-skilled =3	FP	FN	TN

where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative based on the confusion matrix to represent the true prediction of skill level classification across three categories. The formula for performance metrics including specificity, sensitivity, F1-score and accuracy is as below.

Sensitivity measures the number of true positive predictions to the total number of actual positive observations as in Equation 8 [66]. A higher sensitivity score reflects the ability of the skill level prediction model to correctly recognise the three skill level categories. For instance, a high sensitivity value for the skilled category indicates that most of the skilled workers have been accurately assigned to their actual category.

$$Sensitivity = \frac{(TP1+TP2+TP3)}{(TP1+TP2+TP3)+(FN1+FN2+FN3)} \quad (8)$$

Specificity represents the total number of true negative predictions to the number of actual negative observations using Equation 9 [66]. In this study, a high degree of specificity indicates the model's efficiency to accurately excluding workers who do not belong to that particular skill level category. Thus, it minimises misclassification among the three skill level classes.

$$Specificity = \frac{(TN1+TN2+TN3)}{(TN1+TN2+TN3)+(FP1+FP2+FP3)} \quad (9)$$

The F1-score is a single metric combining precision and recall to provide a balanced measure of a model's classification performance as Equation 10 [64]. A high F1-score shows the model's ability to accurately determine workers respective skill level categories while preventing incorrect categorisation.

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

Accuracy refers to the number of true predictions to the total number of observations in the dataset as computed in Equation 11 [66]. It represents the capacity of the classification model to assign workers according to their correct skill level categories. A strong accuracy score indicates that the algorithm has effectively identified worker's actual skill level category.

$$Accuracy = \frac{(TP1+TP2+TP3) + (TN1+TN2+TN3)}{(TP1+TP2+TP3)+ (TN1+TN2+TN3)+(FP1+FP2+FP3)+(FN1+FN2+FN3)} \quad (11)$$

The result of each algorithm was compared to obtain the proposed model for predicting the skill level. Model with optimal result was selected and applied to predict the skill level classification.

Results and Discussion

The analysis of this study can be segmented into three components. It consists of descriptive statistics analysis, performance of tree-based algorithms and feature importance affecting the skill level classification.

Descriptive Statistics

This study utilised the Salaries and Wages Survey for the reference year 2023 with 120,518 cases. The data comprises mixed data types including categorical and continuous features. The distribution of gender was depicted in Figure 5 with male dominating the cases with 62.8%. Meanwhile, Figure 6 illustrates the composition of the economic sector with the majority working in services (64.3%) while fewer labourers work in mining & quarrying (0.7%).

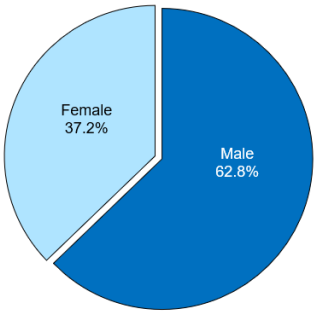

Figure 5. Distribution of Cases by Gender

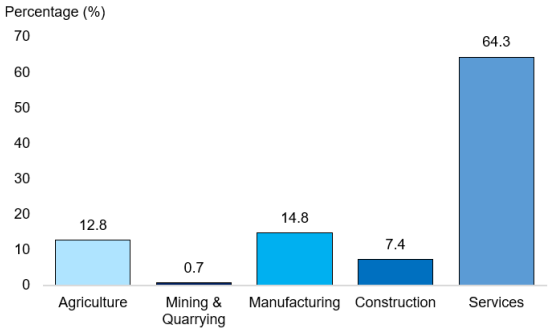

Figure 6. Distribution of Cases by Economic Sector

In terms of the target variable, the skill level can be classified into skilled, semi-skilled and low-skilled labour as depicted in Figure 7. The classes were dominated by the semi-skilled category (74,717) which overrepresented skilled (30,969) and low-skilled (14,832) cases. This indicates the target variables are imbalanced which is an issue for machine learning approaches as it leads to a biased prediction by concentrating on the majority classes and performing poorly within the minority classes [18].

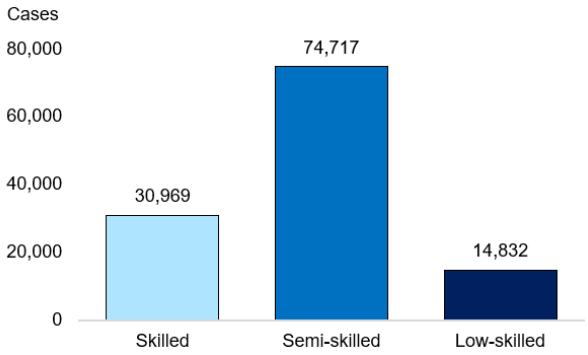

Figure 7. Distribution of Cases by Skill Level

Performance of Tree-Based Algorithms

This section explains the performance of five tree-based algorithms including Decision Tree, Random Forest, Gradient Boosted Trees, Adaptive Boosting and Extreme Gradient Boosting. The measurement is based on the confusion matrix namely specificity, sensitivity, F1-score and accuracy.

Decision Tree

The effectiveness of the Decision Tree algorithm has been evaluated using multiple metrics. The confusion matrix in Table 2 demonstrates the classification outcomes of the Decision Tree in categorising skilled, semi-skilled and low-skilled workers. It allows explanations of the model's ability to accurately identify both positive and negative classes. This model achieved an accuracy of 80.44% in predicting skill level. This indicates the majority of the workers have been allocated to their actual skill level categories with 4,736 skilled, 12,032 semi-skilled and 2,014 low-skilled workers accurately identified. Specificity recorded at 88.50% shows it is capable of identifying cases with negative classes correctly, which indicates a low chance of assigning the workers to incorrect skill level categories. For instance, only 160 low-skilled workers have been misclassified into the skilled category.

Meanwhile, a sensitivity of 80.44% indicates the model's ability to correctly identify workers who belong to their actual skill level classes. It shows that most of the skilled, semi-skilled and low-skilled workers were correctly identified rather than being misclassified. The F1-score of 80.72% reflects an equal measurement between recall and precision in reducing errors on both false positives and false negatives. These outcomes suggest that the Decision Tree provides a consistent baseline algorithm for skill level classification to correctly identify negative cases which is shown by a higher specificity value.

Table 2. Confusion Matrix for Decision Tree Algorithm

		Predicted Class		
		Skilled	**Semi-skilled**	**Low-skilled**
Actual Class	Skilled	4,736	974	188
	Semi-skilled	1,399	12,032	1,158
	Low-skilled	160	689	2,014

Random Forest

The confusion matrix in Table 3 shows the classification performance of the Random Forest algorithm. This model recorded an accuracy of 86.45% with correctly recognised 4,842 skilled, 13,297 semi-skilled and 2,047 low-skilled workers. This suggests that Random Forest demonstrates a strong capability to capture the actual skill level composition in the labour market. The high specificity value of 90.89% emphasises the robustness of Random Forest in correctly identifying negative cases, thus minimising the risk of misclassification across skill level categories. In this scenario, 68 low-skilled workers were misclassified as skilled, 748 low-skilled as semi-skilled and 87 skilled workers were incorrectly labelled as low-skilled workers.

A sensitivity score of 86.45% represents the Random Forest's abilities to determine positive cases efficiently. The results demonstrate that most workers were correctly categorised according to their actual skill level classes. The F1-score of 86.36% highlights a balanced performance within recall and precision. The balance is essential as semi-skilled workers comprised the largest category within the dataset. Overall, Random Forest demonstrates superior performance against the Decision Tree algorithm by achieving higher scores across four evaluation metrics, indicating its accuracy in identifying workers into skilled, semi-skilled and low-skilled categories.

Table 3. Confusion Matrix for Random Forest Algorithm

		Predicted Class		
		Skilled	**Semi-skilled**	**Low-skilled**
Actual Class	Skilled	4,842	969	87
	Semi-skilled	710	13,297	582
	Low-skilled	68	748	2,047

Gradient Boosted Trees

The Gradient Boosted Trees classification performance was evaluated using the confusion matrix shown in Table 4. This algorithm achieved an accuracy of 75.89% in determining the Malaysian workforce into skill level categories. A number of 4,338 skilled workers, 11,425 semi-skilled workers and 1,958 low-skilled workers were accurately classified. Despite the ability to capture correct classification categories, misclassifications were also observed across the skill level categories. In this study, the specificity value of 85.92% indicates the effectiveness of this algorithm in correctly identifying negative cases which reducing the risk of workers being determined to higher or lower skill level categories. The results show that 99 low-skilled workers were misidentified as skilled, 806 low-skilled as semi-skilled and 287 skilled workers were incorrectly labelled as low-skilled workers.

The sensitivity score of 75.89% demonstrates the algorithm's efficiency in accurately determining positive cases in which the workers were correctly identified based on their actual skill level. However, misclassification exists such as 1,281 semi-skilled workers were identified as skilled while 1,883 semi-skilled as low-skilled workers. On the other hand, the F1-score of 76.55% represents the model's efficiency in addressing the class imbalances while maintaining prediction accuracy. Overall, although this algorithm posted lower performance across all evaluation metrics compared to the single tree and other ensemble classifiers, the outcomes remain informative in understanding the classification patterns of workers' skill level.

Table 4. Confusion Matrix for Gradient Boosted Trees Algorithm

		Predicted Class		
		Skilled	**Semi-skilled**	**Low-skilled**
Actual Class	Skilled	4,338	1,273	287
	Semi-skilled	1,281	11,425	1,883
	Low-skilled	99	806	1,958

Adaptive Boosting

The classification performance of Adaptive Boosting to predict skill level is summarised in the confusion matrix as depicted in Table 5. This model achieved an overall accuracy of 68.42% with 4,286 skilled, 9,865 semi-skilled and 1,826 low-skilled workers accurately classified. Despite the algorithm's capacity to capture correct skill level categories, its accuracy score was the lowest compared with the single tree, ensemble and boosting techniques. In terms of specificity, Adaptive Boosting recorded 82.57% showing its capability to determine negative cases correctly while minimising misclassification across skill level categories. For instance, 143 low-skilled workers were labelled as skilled, 894 of the low-skilled workers as semi-skilled and 324 skilled as low-skilled workers.

A sensitivity level of 68.42% reveals weaker performance in determining positive cases. It shows that while this algorithm identified the actual skill level categories, there were also misclassifications among the workers' skill level. The largest misclassification occurred in the semi-skilled category where 2,280 were misclassified as skilled while 2,444 as low-skilled. The F1-score obtained was 69.46% highlighting a balance performance between precision and recall. In conclusion, Adaptive Boosting registered the weakest classification performance among other tree-based algorithms in this study across accuracy, sensitivity, specificity and F1-score. The results indicate the limitations of this algorithm in managing imbalanced labour market data and highlight the need for more robust techniques to ensure classification accuracy across the three skill level classes.

Table 5. Confusion Matrix for Adaptive Boosting Algorithm

		Predicted Class		
		Skilled	**Semi-skilled**	**Low-skilled**
Actual Class	Skilled	4,286	1,288	324
	Semi-skilled	2,280	9,865	2,444
	Low-skilled	143	894	1,826

Extreme Gradient Boosting

The classification performance of Extreme Gradient Boosting (XGBoost) in predicting skill level is presented using the confusion matrix in Table 6. The algorithm obtained a model accuracy of 82.57% which indicates that a large number of workers were correctly assigned into their respective skill level categories. A total of 4,554 skilled, 12,835 semi-skilled and 1,891 low-skilled workers were accurately detected, highlighting the algorithm's capability to accurately determine the actual skill level composition. The specificity of 88.52% indicates its powerful capability to detect the negative cases which lowers the possibilities of misclassifying the workers. As an example, 894 low-skilled workers have been wrongly assigned as semi-skilled workers whereas 130 skilled workers were incorrectly classified as low-skilled workers.

In addition, XGBoost obtained a sensitivity of 82.57% reflecting an excellent level of correctly detecting the positive cases on skill level classification. This shows that the majority of skilled, semi-skilled and low-skilled workers have been successfully identified into their actual skill level classes. The value of the F1-score is 82.53% highlighting a balanced prediction between precision and recall where Extreme Gradient Boosting is reliable for handling class imbalances where semi-skilled workers dominate the dataset. The XGBoost performance across accuracy, sensitivity, specificity and F1-score demonstrate its effectiveness in classifying the skill level among Malaysian workers.

Table 6. Confusion Matrix for Extreme Gradient Boosting Algorithm

		Predicted Class		
		Skilled	**Semi-skilled**	**Low-skilled**
Actual Class	Skilled	4,554	1,214	130
	Semi-skilled	863	12,835	891
	Low-skilled	78	894	1,891

Feature Importance

In Python, feature importance for each algorithm can be analysed according to the weight of each feature. Subsequently, the top 10 features were identified following the feature ranking [67] as relevant features influencing skill level classification [68].

Decision Tree

Figure 8 highlights the top ten relevant features to predict skill level classification. The highest weight was salaries and wages received (0.27) which significantly surpassed the other features. This illustrates that salaries and wages received are the most important in the prediction of skill level classification by using the Decision Tree. This algorithm heavily relies on this feature in the prediction of the target variable. The second highest weight is economic activity (0.17), followed by certificate obtained (0.06) and state of birth (0.05).

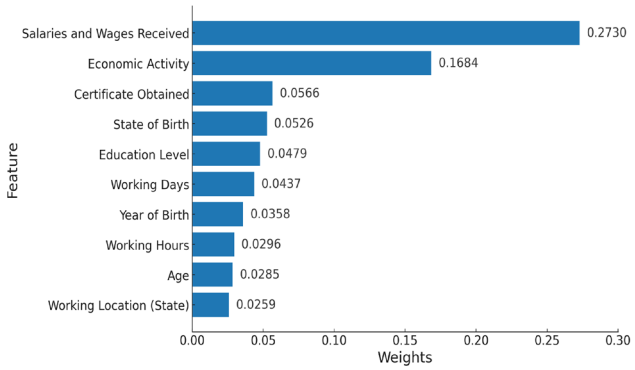

Figure 8. Features Weight Using Decision Tree Algorithm

Random Forest

The weight generated by Random Forest algorithm are depicted in Figure 9. The main feature contribution to predict skill level was salaries and wages received (0.17). Economic activity ranked

second (0.13), followed by education level (0.06) and certificate obtained (0.06). These features show the important features influencing skill level are depending on the higher value of weights.

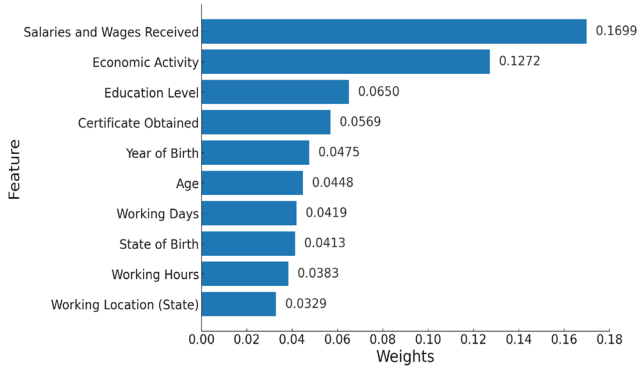

Figure 9. Features Weight Using Random Forest Algorithm

Gradient Boosted Trees

Figure 10 highlights the feature weight generated by Gradient Boosted Trees to provide insights on relevant of input features. Three highest weights were salaries and wages received (0.44), economic activity (0.13) and education level (0.09). These outcomes emphasise the important role of salaries and wages received, economic activity and education level to classify skill level category.

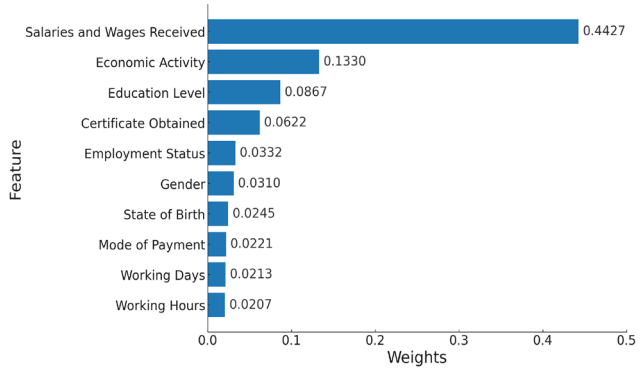

Figure 10. Features Weight Using Gradient Boosted Trees Algorithm

Adaptive Boosting

The top ten features identified based on Adaptive Boosting are illustrated in Figure 11. Economic activity was the most important feature by registering the highest weight of 0.26. Salaries and wages received (0.13) ranked second while working hours (0.10) ranked third. This finding provides insights to the labour market characteristics that influence the skill level classification.

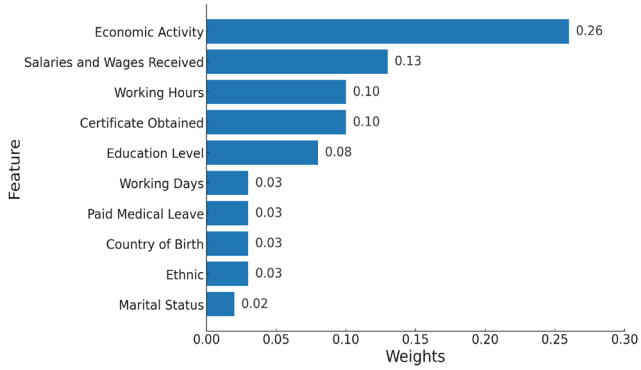

Figure 11. Features Weight Using Adaptive Boosting Algorithm

Extreme Gradient Boosting

There are ten features identified by Extreme Gradient Boosting to predict skill level based on the value of weights as depicted in Figure 12. The three most important features are education level (0.12), salaries and wages received (0.08) and certificate obtained (0.06). The higher weights for these three features indicate its importance as against other features in classifying the skill level.

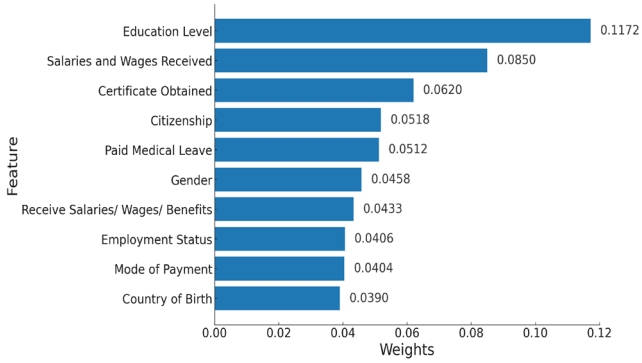

Figure 12. Features Weight Using Extreme Gradient Boosting Algorithm

Performance Comparison

This section emphasises on the evaluation of the overall performance of five tree-based algorithms to classify skill level and the analysis of important features influencing model prediction.

The comparison of five tree-based algorithms including Decision Tree, Random Forest, Gradient Boosted Trees, Adaptive Boosting and Extreme Gradient Boosting is illustrated in Table 7. The result indicates the existence of variations in classification performance across the algorithms. Out of these five algorithms, Random Forest is the most efficient as it achieved the highest value of overall accuracy (86.45%), F1-score (86.36%), sensitivity (86.45%) and specificity (90.89%). According to [69], classification accuracy above 70% is generally considered a strong model performance which supports the efficiency of Random Forest in this study. These findings are in tandem with [37] which have similar performance between predicted and actual classification when employing the Random Forest. A higher F1-score at 86.36% indicates a balanced performance between false positive and false negative which leads to its consistency and reliability [70].

In the meantime, Extreme Gradient Boosting registered the second highest accuracy of 82.57% which was behind Random Forest. This is consistent with an existing study by [71] which similarly reported that Random Forest outperformed Extreme Gradient Boosting. In contrast, Adaptive Boosting recorded the lowest overall accuracy at 68.42% indicating its limitation for multiclass classification studies. The trend is similar across all metrics with the value for specificity, sensitivity and F1-score indicating poor performance among other algorithms. These findings are consistent with [70] which emphasised that Adaptive Boosting is typically less powerful compared to advanced boosting algorithms like Extreme Gradient Boosting.

Table 7. Performance Comparison between Selected Tree-based Algorithms

Model	Specificity (%)	Sensitivity (%)	F1-score (%)	Accuracy (%)
Decision Tree	88.50	80.44	80.72	80.44
Random Forest	90.89	86.45	86.36	86.45
Gradient Boosted Trees	85.92	75.89	76.55	75.89
Adaptive Boosting	82.57	68.42	69.46	68.42
Extreme Gradient Boosting	88.52	82.57	82.53	82.57

In addition of measuring model performance, this study also explored important features affecting skill level classification with each algorithm producing a different set of feature rankings as stated in Table 8. Features including salaries and wages received, economic activity, education level and certificate obtained were frequently identified among the top five important features across the algorithms. This observation aligns with existing studies by [72], [73] and [72], [73] which highlighted salaries and wages

as key indicators in labour market analysis. Furthermore, economic activity was emphasised by [72], [73] and [74] as a relevant feature to determine skill categorisation. Education level has also been widely known as an important factor in predicting skill level classification [73], [75]. Decision Tree and Random Forest recorded the same top ten relevant features but with different feature weights. The differences in feature rankings attributed by the specific mechanisms of each algorithm including splitting method, ensemble procedure and interaction between features.

Table 8. Top 10 Relevant Features between Selected Tree-based Algorithms

No.	Decision Tree	Random Forest	Gradient Boosted Trees	Adaptive Boosting	Extreme Gradient Boosting
1.	Salaries and Wages Received	Salaries and Wages Received	Salaries and Wages Received	Economic Activity	Education Level
2.	Economic Activity	Economic Activity	Economic Activity	Salaries and Wages Received	Salaries and Wages Received
3.	Certificate Obtained	Education Level	Education Level	Working Hours	Certificate Obtained
4.	State of Birth	Certificate Obtained	Certificate Obtained	Certificate Obtained	Citizenship
5.	Education Level	Year of Birth	Employment Status	Education Level	Paid Medical Leave
6.	Working Days	Age	Gender	Working Days	Gender
7.	Year of Birth	Working Days	State of Birth	Paid Medical Leave	Receive Salaries/ Wages/ Benefits
8.	Working Hours	State of Birth	Mode of Payment	Country of Birth	Employment Status
9.	Age	Working Hours	Working Days	Ethnic	Mode of Payment
10.	Working Location (State)	Working Location (State)	Working Hours	Marital Status	Country of Birth

This study employed five supervised tree-based algorithms to predict skill level categories in the Malaysian workforce using household survey data. The findings demonstrate the efficiency of ensemble method particularly the Random Forest which achieved highest classification performance. It shows that Random Forest is able to handle complex and large predictors. Similarly, Extreme Gradient Boosting also provides strong performance by correcting the classification errors from existing models. However, the performance did not surpass Random Forest possibly due to sensitivity of hyperparameter tuning.

In terms of feature relevance, this study identifies several features which play an important role in influencing skill level classification. The salaries and wages received was consistently ranked as a key feature influencing skill level across all algorithms. It highlights that the salary earned by a labour plays a strong role in distinguishing the category of skill. These findings are associated with literature from [72], [73] and [74] as they similarly identified salary as a key factor in their labour market studies. Additionally, several other features that were frequently identified across various algorithms are economic activity, education level and certificate obtained. By leveraging machine learning approaches, this study enables the identification of important features contributing to the classification model. This represents the improvement beyond the existing approaches as it primarily concentrated on occupational groups. The findings provide valuable insights for policymakers to create more targeted training and upskilling interventions aligned with the requirements needed in the labour market.

Conclusions

In conclusion, this study highlights the efficiency of tree-based machine learning algorithms in classifying skill level using household survey data. Among those five algorithms, Random Forest was identified as the proposed model in predicting skill level classification within the Malaysian workforce. It achieved the highest performance with 86.45% of accuracy, 90.89% of specificity and 86.36% of F1-score. This outstanding performance indicates Random Forest capability in managing complex feature relationships, handling imbalanced class distributions and mitigating overfitting although it does not address classification errors during the model training such as boosting techniques. Boosting algorithms such as Extreme Gradient Boosting and single tree algorithm (Decision Tree) also performed well with achieving overall accuracy above 80%. In contrast, Adaptive Boosting categorised under the boosting ensemble method registered the lowest performance with an accuracy of less than 70%. The limitation of this study that needs consideration in future research is the application of a single year dataset which may limit the generalisability of the findings as it does not capture changes in labour market conditions.

Among all algorithms, salaries and wages received, economic activity, certificate obtained and education level were consistently considered as the relevant features influencing skill level classification. It demonstrates important criteria of salary earned by a labourer and education in determining the category of skill level. Other relevant features affecting skill level were working days and working hours. These input features provide important insights which emphasise various factors that contribute the skill classification within the Malaysian labour market across demographic and socio-economic characteristics.

The findings suggest that tree-based algorithms are an alternative in predicting skill level categories by utilising labour market survey data. These algorithms are capable of generating powerful classification performance and are efficient in determining the important input features. In conclusion, tree-based algorithms particularly the ensemble method like Random Forest are powerful in analysing labour market information which can be utilised to support initiatives relating to human capital policy.

Conflicts of Interest

The authors declares that there is no conflict of interest regarding the publication of this paper.

Acknowledgment

Thank you to the Department of Statistics Malaysia for providing the annual labour market data based on the household survey dataset. Appreciation to Universiti Teknologi MARA (UiTM) for providing guidance and academic support throughout the research journey.

References

[1] Economic Planning Unit. (2021). *Twelfth Malaysia Plan 2021–2025*. Putrajaya: Economic Planning Unit.
[2] Gammarano, R. (2019). *Work and employment are not synonyms*. ILOSTAT. https://ilostat.ilo.org/blog/work-and-employment-are-not-synonyms/.
[3] P. J. Boettke, 'Economics in a world amid flux', *Behav. Public Policy*, pp. 1–16, Jan. 2025, doi: 10.1017/bpp.2024.57.
[4] Department of Statistics Malaysia. (2024). *Labour force survey report 2023*. Putrajaya: Department of Statistics Malaysia.
[5] International Labour Organization. (n.d.). *Productivity and skills utilisation*. Skills and Lifelong Learning. https://www.ilo.org/topics-and-sectors/skills-and-lifelong-learning/productivity-and-skills-utilisationand-skills-utilisation.
[6] Ministry of Human Resources. (2020). *Malaysia standard classification of occupations 2020*. Putrajaya: Ministry of Human Resources.
[7] Bukola, O. A., & Tosin. (2023). Introduction to descriptive statistics. In *Recent advances in biostatistics*. IntechOpen.
[8] Hashim, N. M., Noor, N. M., Ul-Saufie, A. Z., Sandu, A. V., Vizureanu, P., Deák, G., & Kheimi, M. (2022). Forecasting daytime ground-level ozone concentration in urbanized areas of Malaysia using predictive models. *Sustainability, 14*(13), 7936. https://doi.org/10.3390/su14137936.
[9] Asadi, F., Homayounfar, R., Mehrali, Y., Masci, C., & Talebi, S. (2024). Detection of cardiovascular disease cases using advanced tree-based machine learning algorithms. *Scientific Reports, 14*(1), 22230. https://doi.org/10.1038/s41598-024-72819.
[10] Ribeiro Junior, R. F., & Gomes, G. F. (2024). On the use of machine learning for damage assessment in composite structures: A review. *Applied Composite Materials, 31*(1), 1–37. https://doi.org/10.1007/s10443-023-10161-5.
[11] Anuar, A., Mohd Hussain, N. H., & Byrd, H. (2023). Tree-based machine learning in classifying reverse migration. *Mathematical Sciences and Informatics Journal*, *4*(1), 49–56.
[12] Baghdadi, A., Lama, S., Singh, R., & Sutherland, G. R. (2023). Tool-tissue force segmentation and pattern recognition for evaluating neurosurgical performance. *Scientific Reports, 13*(1), 9591. https://doi.org/10.1038/s41598-023-36702-3.
[13] Dials, J., *et al.* (2023). Skill-level classification and performance evaluation for endoscopic sleeve gastroplasty. *Surgical Endoscopy, 37*(6), 4754–4765. https://doi.org/10.1007/s00464-023-09955-2
[14] Soleymani, A., Sadat Asl, A. A., Yeganejou, M., Dick, S., Tavakoli, M., & Li, X. (2021). Surgical skill evaluation from robot-assisted surgery recordings. In *2021 International Symposium on Medical Robotics (ISMR)* (pp. 1–6). IEEE. https://doi.org/10.1109/ISMR48346.2021.9661527.
[15] Chen, M., Hui Fang Szu, Hsin Yen Lin, Liu, Y., Ho Yin Chan, Wang, Y., Zhao, Y., Zhang, G., Yao, D., & Li, W. J. (2023). Phase-based quantification of sports performance metrics using a smart IoT sensor. *IEEE Internet of Things Journal, 10*(18), 15900–15911. https://doi.org/10.1109/JIOT.2023.3266351.
[16] Guo, X., Brown, E., Chan, P. P. K., Rosa, & Cheung, R. T. H. (2023). Skill level classification in basketball free-throws using a single inertial sensor. *Applied Sciences, 13*(9), 5401. https://doi.org/10.3390/app13095401.

[17] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953.
[18] Sayed, E. H., Alabrah, A., Rahouma, K. H., Zohaib, M., & Badry, R. M. (2024). Machine learning and deep learning for loan prediction in banking: Exploring ensemble methods and data balancing. *IEEE Access, 12*, 193997–194019. https://doi.org/10.1109/ACCESS.2024.3509774.
[19] Department of Statistics Malaysia. (2024). *Salaries & wages survey 2023*. Putrajaya: Department of Statistics Malaysia.
[20] Dou, J., Song, Y., Wei, G., & Zhang, Y. (2022). Fuzzy information decomposition incorporated and weighted Relief-F feature selection: When imbalanced data meet incompletion. *Information Sciences, 584*, 417–432. https://doi.org/10.1016/j.ins.2021.10.057.
[21] Ul-Saufie, A. Z., Hamzan, N. H., Zahari, Z., Shaziayani, W. N., Noor, N. M., Zainol, M. R. R. M. A., Sandu, A. V., Deak, G., & Vizureanu, P. (2022). Improving air pollution prediction modelling using wrapper feature selection. *Sustainability, 14*(18), 11403. https://doi.org/10.3390/su141811403.
[22] Plante, J.-F., & Radatz, M. (2024). On the capability of classification trees and random forests to estimate probabilities. *Journal of Statistical Theory and Practice, 18*(2), 25. https://doi.org/10.1007/s42519-024-00376-5.
[23] Fellini, I., & Megna, F. (2024). Labour market participation of second-generation youth in Italy. *Rivista Italiana di Economia, Demografia e Statistica*, 78(3), 147–158. https://doi.org/10.71014/sieds.v78i3.289.
[24] Ernas, S. (2024). Over-education rates and predictors of entry-level jobs in Türkiye. *International Journal of Assessment Tools in Education, 11*(4), 758–773. https://doi.org/10.21449/ijate.1495346.
[25] Bischof, S. (2024). Test-based measurement of skill mismatch: A validation of five different measurement approaches using the NEPS. *Journal for Labour Market Research, 58*(1), 11. https://doi.org/10.1186/s12651-024-00370-1.
[26] Van Oosten, A. J., Van Mens, K., Blonk, R. W. B., Burdorf, A., & Tiemens, B. (2023). The relationship between having a job and the outcome of brief therapy in patients with common mental disorders. *BMC Psychiatry, 23*(1), 910. https://doi.org/10.1186/s12888-023-05418-z.
[27] Kiss, Z. (2024). Vertical and horizontal (mis)match of university degrees in the Hungarian labour market. *Humanities and Social Sciences Communications, 11*(1), 1699. https://doi.org/10.1057/s41599-024-04203-x.
[28] Gultekin, D., Hisarciklilar, M., & Yusufi, F. (2024). Multiple faces of labour market segmentation within the Turkish construction industry. *Economic and Labour Relations Review*, 1–22. https://doi.org/10.1017/elr.2024.35.
[29] Omia, E., Bae, H., Park, E., Kim, M. S., Baek, I., Kabenge, I., & Cho, B.-K. (2023). Remote sensing in field crop monitoring: A comprehensive review of sensor systems, data analyses and recent advances. *Remote Sensing, 15*(2), 354. https://doi.org/10.3390/rs15020354.
[30] Celbiş, M. G., Wong, P., Kourtit, K., & Nijkamp, P. (2023). Impacts of the COVID-19 outbreak on older-age cohorts in European labor markets: A machine learning exploration of vulnerable groups. *Regional Science Policy & Practice, 15*(3), 559–585. https://doi.org/10.1111/rsp3.12520.
[31] Mansor, A., & Othman, Z. (2024). Malaysian community college graduates employability prediction model using machine learning approach. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 16*(3), 1–7. https://doi.org/10.54554/jtec.2024.16.03.001.
[32] Ong, S. Y., Ting, C. Y., Goh, H. N., Quek, A., & Cham, C. L. (2023). Workplace preference analytics among graduates. *Journal of Informatics and Web Engineering, 2*(2), 233–248. https://doi.org/10.33093/jiwe.2023.2.2.17.
[33] Lee, J.-Y., Lee, W., & Cho, S. (2023). Characteristics of fatal occupational injuries in migrant workers in South Korea: A machine learning study. *Heliyon, 9*(9), e20138. https://doi.org/10.1016/j.heliyon.2023.e20138.
[34] Cho, W. H., Shin, J., Kim, Y. D., & Jung, G. J. (2022). Pixel-wise classification in graphene detection with tree-based machine learning algorithms. *Machine Learning: Science and Technology, 3*(4), 045029. https://doi.org/10.1088/2632-2153/aca744.
[35] Lin, W.-C., Huang, C.-H., Chien, L.-T., Tseng, H.-J., Ng, C.-J., Hsu, K.-H., Lin, C.-C., & Chien, C.-Y. (2022). Tree-based algorithms and association rule mining for predicting patients' neurological outcomes after first-aid treatment for an out-of-hospital cardiac arrest during the COVID-19 pandemic: Application of data mining. *International Journal of General Medicine, 15*, 7395–7405. https://doi.org/10.2147/IJGM.S384959.
[36] Mansur, R., & Subroto, A. (2022). Using tree-based algorithm to predict informal workers' willingness to pay national health insurance after tele-collection. *2022 10th International Conference on Information and Communication Technology (ICoICT)*, 23–28. https://doi.org/10.1109/ICoICT55009.2022.9914901.
[37] Shaziayani, W. N., Ul-Saufie, A. Z., Mutalib, S., Mohamad Noor, N., & Zainordin, N. S. (2022). Classification prediction of PM10 concentration using a tree-based machine learning approach. *Atmosphere*, *13*(4), 538. https://doi.org/10.3390/atmos13040538.
[38] Soangra, R., Sivakumar, R., Anirudh, E. R., Reddy, S. V., & John, E. B. (2022). Evaluation of surgical skill using machine learning with optimal wearable sensor locations. *PLOS ONE, 17*(6), e0267936. https://doi.org/10.1371/journal.pone.0267936.
[39] Malek, N. H. A., Wan Yaacob, W. F., Md Nasir, S. A., & Shaadan, N. (2022). Prediction of water quality classification of the Kelantan River Basin, Malaysia, using machine learning techniques. *Water, 14*(7), 1067. https://doi.org/10.3390/w14071067.
[40] Toharudin, T., *et al.* (2023). Boosting algorithm to handle unbalanced classification of PM2.5 concentration levels by observing meteorological parameters in Jakarta, Indonesia using AdaBoost, XGBoost, CatBoost, and LightGBM. *IEEE Access, 11*, 35680–35696. https://doi.org/10.1109/ACCESS.2023.3265019.
[41] S., Hu, Y., Zhang, L., Liu, S., Xie, R., & Yin, Z. (2024). Intelligent risk identification for drilling lost circulation incidents using data-driven machine learning. *Reliability Engineering & System Safety, 252*, 110407. https://doi.org/10.1016/j.ress.2024.110407.
[42] International Labour Office. (1979). *An integrated system of wages statistics*. Geneva: International Labour Office.

[43] Salah-Ud-Din, M., B. T. L. S. S., & Al Ali, H. (2024, April). Exploratory data analysis and prediction of passenger satisfaction with airline services. In *2024 New Trends in Civil Aviation (NTCA)* (pp. 295–302). IEEE. https://doi.org/10.23919/NTCA60572.2024.10517814.
[44] Shin, H., & Lee, S. (2021). An OMOP-CDM-based pharmacovigilance data-processing pipeline (PDP) providing active surveillance for ADR signal detection from real-world data sources. *BMC Medical Informatics and Decision Making, 21*(1), 159. https://doi.org/10.1186/s12911-021-01520-y.
[45] Elmannai, H., Alqahtani, A., Mahfoud, A., Khan, R. A., Alotaibi, S. S., & Alghamdi, A. (2023). Polycystic ovary syndrome detection machine learning model based on optimized feature selection and explainable artificial intelligence. *Diagnostics, 13*(8), 1506. https://doi.org/10.3390/diagnostics13081506.
[46] Kim, T.-Y., Lee, S., Park, H., Kim, Y., Lee, S., & Lim, J. (2024). Occupation classification model based on DistilKoBERT: Using the 5th and 6th Korean Working Condition Surveys. *Annals of Occupational and Environmental Medicine, 36*(1), e19. https://doi.org/10.35371/aoem.2024.36.e19
[47] Baptiste, P. J., Wong, A. Y. S., Schultze, A., Clase, C. M., Clémence Leyrat, Williamson, E., Powell, E., Mann, J. F. E., Cunnington, M., Teo, K., Bangdiwala, S. I., Gao, P., Wing, K., & Tomlinson, L. (2024). Effectiveness and risk of ARB and ACEi among different ethnic groups in England: A reference trial (ONTARGET) emulation analysis using UK Clinical Practice Research Datalink Aurum-linked data. *PLOS Medicine, 21*(9), e1004465. https://doi.org/10.1371/journal.pmed.1004465.
[48] Zolbanin, H., & Aubert, B. (2025). A process model for design-oriented machine learning research in information systems. *Journal of Strategic Information Systems, 34*(1), 101868. https://doi.org/10.1016/j.jsis.2024.101868.
[49] Lartey, C., Liu, J., Asamoah, R. K., Greet, C., Zanin, M., & Skinner, W. (2024). Effective outlier detection for ensuring data quality in flotation data modelling using machine learning (ML) algorithms. *Minerals, 14*(9), 925. https://doi.org/10.3390/min14090925.
[50] Wang, J., Ueda, T., Wang, P., Li, Z., & Li, Y. (2025). Building damage inspection method using UAV-based data acquisition and deep learning-based crack detection. *Journal of Civil Structural Health Monitoring, 15*(1), 151–171. https://doi.org/10.1007/s13349-024-00836-3.
[51] Nasaruddin, N., Masseran, N., Idris, W. M. R., & Ul-Saufie, A. Z. (2024). Reduced noise SMOTE in machine learning model: Application in water quality classification with imbalanced datasets. In *2024 5th International Conference on Artificial Intelligence and Data Sciences (AiDAS)* (pp. 87–92). IEEE. https://doi.org/10.1109/AiDAS63860.2024.10730391.
[52] Shaha, T. R., Begum, M., Uddin, J., Torres, V. Y., Iturriaga, J. A., Ashraf, I., & Samad, M. A. (2024). Feature group partitioning: An approach for depression severity prediction with class balancing using machine learning algorithms. *BMC Medical Research Methodology, 24*(1), 123. https://doi.org/10.1186/s12874-024-02249-8
[53] Unlu, A., & Subasi, A. (2025). Substance use prediction using artificial intelligence techniques. *Journal of Computational Social Science, 8*(1), 21. https://doi.org/10.1007/s42001-024-00356-6.
[54] Uddin, S., & Lu, H. (2024). Confirming the statistically significant superiority of tree-based machine learning algorithms over their counterparts for tabular data. *PLOS ONE, 19*(4), e0301541. https://doi.org/10.1371/journal.pone.0301541.
[55] Mustakim, N. A., Ul-Saufie, A. Z., Shaziayani, W. N., Mohamad Noor, N., & Mutalib, S. (2022). Prediction of daily air pollutants concentration and air pollutant index using machine learning approach. *Pertanika Journal of Science & Technology, 31*(1), 123–135. https://doi.org/10.47836/pjst.31.1.08.
[56] Alharbi, A. A. (2024). Classification performance analysis of decision tree-based algorithms with noisy class variable. *Discrete Dynamics in Nature and Society, 2024*, 6671395. https://doi.org/10.1155/2024/6671395.
[57] Rezaei, A., Yazdinejad, M., & Sookhak, M. (2024). Credit card fraud detection using tree-based algorithms for highly imbalanced data. In *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICMI60790.2024.10586088.
[58] Imada, J., Arango-Sabogal, J. C., Bauman, C., Roche, S., & Kelton, D. (2024). Comparison of machine learning tree-based algorithms to predict future paratuberculosis ELISA results using repeat milk tests. *Animals, 14*(7), 1113. https://doi.org/10.3390/ani14071113.
[59] Ibrahim, S., Balzter, H., & Tansey, K. (2024). Machine learning feature importance selection for predicting aboveground biomass in African savannah with Landsat 8 and ALOS PALSAR data. *Machine Learning with Applications, 16*, 100561. https://doi.org/10.1016/j.mlwa.2024.100561.
[60] Islam, M. K., Reza, I., Gazder, U., Akter, R., Arifuzzaman, M., & Rahman, M. M. (2022). Predicting road crash severity using classifier models and crash hotspots. *Applied Sciences, 12*(22), 11354. https://doi.org/10.3390/app122211354.
[61] Mienye, I. D., & Jere, N. (2024). A survey of decision trees: Concepts, algorithms, and applications. *IEEE Access, 12*, 86716–86727. https://doi.org/10.1109/ACCESS.2024.3416838.
[62] Champahom, T., Se, C., Watcharamaisakul, F., Jomnonkwao, S., Karoonsoontawong, A., & Ratanavaraha, V. (2024). Tree-based approaches to understanding factors influencing crash severity across roadway classes: A Thailand case study. *IATSS Research, 48*(3), 464–476. https://doi.org/10.1016/j.iatssr.2024.09.001.
[63] Putra, M., Rosid, M. S., & Handoko, D. (2022). Rainfall estimation using machine learning approaches with raingauge, radar, and satellite data. In *2022 International Conference on Electrical Engineering and Informatics (ICELTICs)* (pp. 25–30). IEEE. https://doi.org/10.1109/ICELTICs56128.2022.9932109.
[64] Asteris, P. G., Rizal, F. I. M., Koopialipoor, M., Roussis, P. C., Ferentinou, M., Armaghani, D. J., & Gordan, B. (2022). Slope stability classification under seismic conditions using several tree-based intelligent techniques. *Applied Sciences, 12*(3), 1753. https://doi.org/10.3390/app12031753
[65] Wang, Z., He, C., Hu, Y., Luo, H., Li, C., Wu, X., Zhang, Y., Li, J., & Cai, J. (2024). A hybrid deep learning scheme for MRI-based preliminary multiclassification diagnosis of primary brain tumors. *Frontiers in Oncology, 14*, 1363756. https://doi.org/10.3389/fonc.2024.1363756.
[66] Ibrahim, N., Ishak, U. M., Ali, N. N. A., Shaadan, N., & others. (2024). Machine learning-based approaches for credit card debt prediction. *Malaysian Journal of Computing, 9*(1), 1722–1733. https://doi.org/10.24191/mjoc.v9i1.25656.

[67] Javeed, A., Anderberg, P., Ghazi, A. N., Saleem, M. A., & Sanmartin Berglund, J. (2025). Predicting depression in older adults: A novel feature selection and neural network framework. *Neural Processing Letters, 57*(3), 41. https://doi.org/10.1007/s11063-025-11760-y.
[68] Hu, P., & Zhu, J. (2025). A filter-wrapper model for high-dimensional feature selection based on evolutionary computation. *Applied Intelligence, 55*(7), 581. https://doi.org/10.1007/s10489-025-06474-6.
[69] Halias, A. F., Saiful, N. H., Ibrahim, N., Muhamad Jamil, S. A., Mansor, M. M., Ul - Saufie, A. Z., & Md Ghani, N. A. (2023). Type 2 diabetes mellitus prediction using data mining approach. *2023 IEEE International Conference on Computing (ICOCO), 2824*, 29–34. https://doi.org/10.1109/icoco59262.2023.10398078.
[70] Lawal, Z. K., Aldrees, A., Yassin, H., Salisu Dan'azumi, Sujay Raghavendra Naganna, Abba, S. I., & Saad Sh. Sammen. (2024). Optimized ensemble methods for classifying imbalanced water quality index data. *IEEE Access*, 1–1. https://doi.org/10.1109/access.2024.3502361.
[71] Mehmood, K., Shoaib Ahmad Anees, Luo, M., Akram, M., Zubair, M., Khan, K. A., & Khan, W. R. (2024). Assessing Chilgoza Pine (*Pinus gerardiana*) forest fire severity: Remote sensing analysis, correlations, and predictive modeling for enhanced management strategies. *Trees, Forests and People, 16*, 100521. https://doi.org/10.1016/j.tfp.2024.100521.
[72] Afsharinia, B., & Gurtoo, A. (2024). COVID-19 impact on food consumption of low-skilled employees in India. *Global Food Security, 42*, 100791. https://doi.org/10.1016/j.gfs.2024.100791.
[73] Josten, C., Krause, H., Lordan, G., & Yeung, B. (2024). What skills pay more? The changing demand and return to skills for professional workers. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4706059.
[74] Yang, G., Yao, S., & Dong, X. (2023). Digital economy and wage gap between high- and low-skilled workers. *Digital Economy and Sustainable Development, 1*(1), 7. https://doi.org/10.1007/s44265-023-00009-y.
[75] Kaboth, A., Hünefeld, L., & Lück, M. (2024). Exploring work ability, psychosocial job demands and resources of employees in low-skilled jobs: A German cross-sectional study. *Journal of Occupational Medicine and Toxicology, 19*(1), 30. https://doi.org/10.1186/s12995-024-00429-2.