E-ISSN 2586-9027
Homepage : https://tci-thaijo.org/index.php/SciTechAsia Science & Technology Asia
Vol.31 No.1 January - March 2026 Page: [145-158]
*Original research article*

# Data Mining Model Approach for Employment Prediction for University Graduates

Tewa Promnuchanont, Sureenat Manola*, Worakarn Jaidee

*Department of Business Information System, Faculty of Business Administration and Liberal Arts, Rajamangala University of Technology Lanna, Chiang Mai 50300, Thailand*

Received 24 June 2025; Received in revised form 3 December 2025
Accepted 7 January 2026; Available online 27 March 2026

**ABSTRACT**

Graduate employability prediction has become increasingly important as universities seek to understand the factors influencing employment outcomes and to improve academic planning. However, prior studies in this area often rely on limited datasets, evaluate only a narrow range of models, and lack systematic feature assessment, which restricts the robustness and generalizability of their findings. This study addresses these limitations by developing a comprehensive multi-model data-mining framework to predict the employability of graduates from Rajamangala University of Technology Lanna (RMUTL). A dataset of 4,352 graduate records from the 2023 academic year was analyzed. Three filter-based feature-selection techniques—chi-square, information gain, and correlation-based evaluation—were applied to identify the most influential predictors. Five machine-learning algorithms (Decision Tree, Random Forest, Gradient Boosted Trees, Naïve Bayes, and K-Nearest Neighbors) were trained and evaluated using accuracy, precision, recall, F1-score, and AUC. The results show that Random Forest achieved the highest accuracy (83.72%), while Gradient Boosted Trees yielded the highest AUC (0.813), indicating superior class-separation performance. Key predictive factors identified across models included curriculum, education level, department, faculty, campus, gender, and GPA level. This study provides a structured comparative modeling framework and identifies institution-specific predictors that influence graduate employability. The findings offer practical implications for curriculum enhancement, evidence-based academic planning, and career-guidance development aimed at improving employment outcomes for RMUTL graduates.

**Keywords:** Data mining; Employability; Evaluate model performance; Forecast

***Corresponding author:** sureenat@rmutl.ac.th **doi:** xx.xxxxx/scitechasia.xxxx.xx

## 1. Introduction

The Ministry of Higher Education, Science, Research, and Innovation in Thailand oversees numerous higher education institutions, both public and private. Among these is Rajamangala University of Technology Lanna (RMUTL), which operates in alignment with its institutional vision. The university's mission emphasizes vocational and technological education as well as the development of professional educators at national and international levels, with the overarching goal of producing graduates who possess practical competencies and meet the quality and professionalism standards of the labor market. Given this mission, graduate employment outcomes have become an increasingly important measure of institutional effectiveness, highlighting the need for RMUTL to better understand and support the employability of its graduates.

One of the major challenges within Thai higher education is the persistent imbalance between the number of graduates produced and the actual demand for skilled labor. This mismatch contributes to underemployment and highlights the need for more effective workforce planning. Although this issue has been widely discussed at the national level, empirical evidence at the institutional level—especially in vocational and technology-oriented universities such as RMUTL—remains limited. As a result, universities lack institution-specific insights that could help identify at-risk student groups, adjust curricula, or design targeted interventions. This represents a clear gap between national policy discussions and institution-level data-driven practice.

Despite RMUTL's initiatives to support students, the university currently lacks a systematic, predictive framework to analyze employment outcomes. Existing graduate data remains an underutilized asset, often stored passively rather than being used for strategic forecasting. In the era of big data, Artificial Intelligence (AI) and Data Mining have become essential tools for decision-making across various industries. Previous studies have demonstrated the efficacy of machine-learning models—such as Decision Trees [1], Random Forests [2], Gradient Boosted Trees [3], Naïve Bayes [4], and K-Nearest Neighbor [5]—in diverse domains ranging from groundwater assessment [11] to genomic classification [12]. However, the application of these advanced techniques within the context of Thai vocational higher education remains limited. Specifically, no prior research has systematically applied these predictive models to RMUTL's unique institutional dataset to forecast employability, representing a significant research gap.

To address this gap, this research aims to develop and evaluate predictive models for RMUTL graduate employability. The novelty of this study is twofold. First, unlike general educational studies, this research integrates specific institutional variables unique to a multi-campus vocational university—including curriculum type, faculty, graduate level, campus location, GPA, and job status—to capture the complex factors influencing technical employment. Second, the study employs a rigorous comparative analysis of multiple data mining techniques (Gradient Boosted Trees, Random Forests, K-Nearest Neighbors, Decision Trees, and Naïve Bayes) combined with advanced feature-selection methods. This approach goes beyond simple descriptive statistics, offering a robust, algorithmic framework to identify the most significant predictors of employability.

The objective of this research is to identify key characteristics influencing em-

ployment outcomes and to determine the most accurate predictive model for the university. The findings will provide actionable insights for RMUTL administrators to enhance curriculum development, design targeted career guidance services, and improve strategic workforce planning. The methodology follows standard data preparation and model development procedures, utilizing RapidMiner Version 10.1 to analyze integrated institutional databases. The remainder of this paper is organized as follows: Section 2 reviews relevant literature, Section 3 details the research methodology, Section 4 presents the results and discussion, and Section 5 concludes the study.

## 2. Materials and Methods

The main program of the proposed algorithm is shown in Fig. 1.

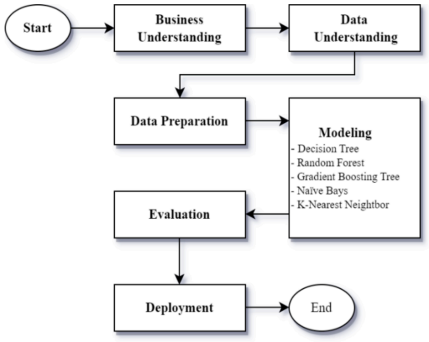



**Fig. 1.** Main program of the proposed algorithm.

The primary program of the proposed algorithm, as depicted in Fig. 1, is structured by the CRISP-DM methodology [13]. The iterative nature of CRISP-DM distinguishes it from other methodologies, enabling adaptability and flexibility tailored to specific project requirements. This iterative approach encompasses six key steps: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. These phases are interlinked and iterative throughout the project lifecycle, underscoring the importance of feedback loops for refining and enhancing outcomes.

### 2.1 Business understanding

The context for this study revolves around the landscape of higher education in Thailand, characterized by a proliferation of public and private institutions. Despite this, there exists an imbalance between the supply of graduates and the demands of the job market, posing a significant barrier to the nation's advancement. Many graduates find themselves ill-prepared for the workforce, leading to challenges in securing employment matching their qualifications. To address this issue, our study aims to predict graduates' employability by examining factors influencing their job prospects and developing a predictive model.

### 2.2 Data understanding

The prediction of graduates' employability hinges on the collection and analysis of relevant data. We leverage datasets comprising graduate employability records and students' academic performance averages across various faculties. Data from the academic year are utilized, sourced from institutions such as the Faculty of Engineering, Faculty of Agricultural Science and Technology, Faculty of Fine Arts and Architecture, Faculty of Business Administration and Liberal Arts, and the College of Technology and Interdisciplinary Studies. These datasets are compiled in Microsoft Excel spreadsheets provided by university districts, totaling 4,352 records.

### 2.3 Data preparation

We performed data cleaning, which involved removing data from records that

contained inaccurate, redundant, or both types of information. It is necessary to do an analysis on a total of 3,561 pieces of data at this point in time. Developing a model for the goal of projecting graduate employability was accomplished through the utilization of the attribute selection approach. We refer to the process of selecting features as the feature selection process. This is a method that allows us to select valuable qualities and contribute to model simplification. We use this approach to examine characteristics or factors that influence the prediction of employment opportunities for graduates. It is of the utmost necessity to take attributes into consideration when it comes to generating models that are both more accurate and efficient. At this stage of the process, we apply both domain-knowledge-based and filter-based approaches. Through the application of domain knowledge approaches, the selection of qualities that have a direct impact on the prediction outcomes and the elimination of irrelevant attributes are accomplished. This was followed by the application of three methods: the Chi-Square statistics, information gain, and correlation-based analysis.

I. Chi-square statistics [14] is a filter-based feature selection strategy that uses statistical tests to screen attributes. The Chi-square statistic measures the strength of association between an attribute and the class label and is used to rank attributes according to their statistical importance, as shown in Eq. (2.1).

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \qquad (2.1)$$

where $x^2$ represents Chi-Square statistics, $O_i$ represents observed frequency, and $E_i$ represents expected frequency, which is equal to the quantity of data multiplied by the proportion predicted to be present.

II. The information gain [15] is a feature selection strategy that is utilized to evaluate the worth of data division by computing the gain value for each dimension. This technique is included in the information selection technique. If there is a dimension that possesses the highest gain value, then that dimension is selected to be a subgroup which possesses the classification power of Eq. (2.2).

$$Gain = Entropy(p) - \sum_{i=1}^{k} (\frac{n_i}{n} Entropy(i)), \qquad (2.2)$$

where $Entropy(p)$ is the entropy of the root node before the split, $k$ is the number of child nodes generated by splitting on a given attribute, $n_i$ is the number of samples in child node $i$, $n$ is the total number of samples in the parent node, and $Entropy(i)$ is the entropy of child node $i$ computed from Eq. (2.3).

$$Entropy(i) = -\sum_{j=1}^{m} p(j|i) \log_2 p(j|i), \qquad (2.3)$$

where $p(j|i)$ is the proportion of samples in node $i$ that belong to class $j$, $m$ is the total number of classes, and the conditional probabilities satisfy the normalization condition $\sum_{j=1}^{m} p(j|i) = 1$.

III. The correlation-based methodology [16] is a method for choosing attribute qualities that consider the relationships between attribute groups that are determined by valuing predictability. It can also be used to manage unimportant attributes by selecting and ranking subgroups of dimensions. Information that has no association

with other classes and is substantially correlated with a particular class as Eq. (2.4).

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \qquad (2.4)$$

where $\overline{r_{cf}}$ is the average association between dimensions, $\overline{r_{ff}}$ is the mean relationship of the variable to class ($f \in s$), and $M_s$ is the searchable value of the $S$ subgroup dimension, which contains dimension $k$.

The characteristic with the greatest weight value will be produced from the eight attributes with the weight value as indicated in Table 1 by employing three different ways to determine the weight value and rank the weight value.

Table 1 shows the attribute sequencing based on statistical significance. An attribute's maximum weight will be determined by combining the following 7 variables: curriculum, edu_level, department, faculty, area, gender, and gpa_level.

**Table 1.** Attribute sequencing based on statistical significance.

| Attribute | Chi-Square | Information Gain | Correlation |
|---|---|---|---|
| curriculum | 1135.190 | 0.237 | 0.102 |
| edu_level | 953.667 | 0.201 | 0.460 |
| department | 568.270 | 0.109 | 0.069 |
| faculty | 323.005 | 0.058 | 0.077 |
| area | 93.885 | 0.016 | 0.073 |
| gender | 32.268 | 0.005 | 0.086 |
| gpa_level | 27.061 | 0.004 | 0.004 |

The fact that the data is both numerical and alphabetic is the reasoning behind this. In order to collect data in a format that is suitable for analysis in line with data mining techniques, we have changed the data by exchanging the values of the data. This was done in order to gain suitable data. A person's gender, grade point average, level of graduation, and work position are the three characteristics that are considered. Listed below are the particulars of the modification that has been made available. Statistical information relative to gender the underlying data did not indicate gender, so the data were transformed based on titles. For example, the prefix "Mr." was changed to "male." This was done because the data did not define gender. We have modified the data structure for the average academic performance (GPA) data by dividing the GPA levels into five distinct levels. This was done in order to maintain the integrity of the data. Excellent (GPA > 3.5) to Fail (GPA < 2.0) ratings provide a clear marker for academic accomplishment. Those with a GPA of 3.5 or more are recognized for their academic achievements, while those below 2.0 are identified for further assistance. There are now only two outcome data points for the employment status data, which was the original result data. Regarding this job status, there were initially eight findings in total. The information related to the picture status has been altered to "yes", in contrast to the image status, which has remained at "no". A snippet will be generated after the complete data change process has been successfully completed. The employment status data (job status), which was the result data, was reduced from eight results to just two outcome data. The status of the unemployed is changed to No, and the status of the employed is changed to Yes. The process of modifying and categorizing data, as exemplified by the structured datasets provided, enhances the clarity and usefulness of the information for analysis and decision-making. By organizing attributes such as Gender, Curriculum, Field of Study, Faculty, Graduate Level, Campus, GPA Level, and Job Status into a standardized format, stakeholders can more effectively interpret and utilize the data.

### 2.4 Modeling

The data analysis at this stage involves the utilization of models and data mining techniques, encompassing five distinct procedures: Decision Tree, Random Forest, Gradient Boosting Tree, Naïve Bayes, and K-Nearest Neighbor analysis.

The machine learning model known as a decision tree (DT) is multifunctional and user-friendly, and it may be utilized for classification as well as regression problems. In terms of its conceptual representation, it is comparable to a flowchart, with each internal node representing a feature (or attribute) and a choice depending on the value of that feature, which then leads to following branches that reflect the various possible outcomes. An ultimate choice or forecast is represented by the leaf nodes of the tree. Decision tree prediction categorizes occurrences that qualify for the underlying event. Each hierarchy makes decisions hierarchically. It has 3 symbols: 1) The root node is the uppermost intermediate node that clearly represents the decision norm; 2) the line represents the decision link; and 3) the bottom node, called the leaf node, represents the group of data that will be forecast to the root node of the decision tree, or the decision results. The decision tree process must choose the optimum properties to lay out the root nodes for separation. Eq. (1) shows the number of branches, sub-branches, and other branches with feasibility values. Fig. 2 shows each variable (C1, C2, and C3) as a circle, and the judgment outcomes (Class A and Class B) as rectangles. To properly categorize a sample to a class, each branch is labeled as ”True” or ”False” based on the outcome of the test of its preceding node.

According to the Random Forest approach (RF), in order to acquire the desired results, it is necessary to execute the same

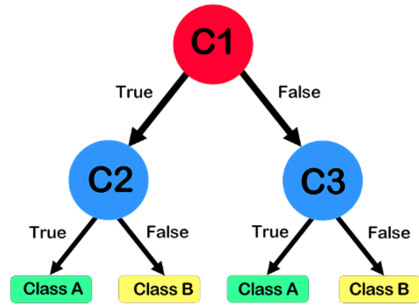



**Fig. 2.** A decision tree (DT) illustration.

predictive model several times on the same data set. Each time a training session is carried out, a distinct subset of the training data is selected. The decisions that are produced by the prediction models are then used to cast a vote on which class is selected the majority of the time. As a consequence of this, the problem of excessive variance that originated with each tree would be resolved. After selecting the master from several different decision trees, with the number of predictive models ranging from ten to more than one thousand, each of which receives a distinct collection of data sets and generates its own forecasts, the final prediction result is selected from the prediction values that received the most votes. A few different decision trees are selected. Fig. 3 shows an example of a random forest made up of three distinct decision trees. A random subset of the training data was used to train each of those three decision trees.

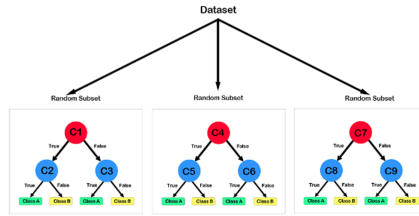



**Fig. 3.** Random forest approach (RF) illustration.

The Gradient Boosting Tree technique (GBT) has gained prominence due to its superior performance in predictive modelling tasks. This technique is a powerful ensemble learning technique. Gradient Boosting Tree is a method that produces a sequence of decision trees progressively, with each tree learning to remedy the flaws of its predecessors. This contrasts with Random Forest, which groups together numerous decision trees at the same time. Gradient descent optimization is the foundation upon which GBT is built. In this approach, the method minimizes a loss function by iteratively fitting new trees to the residual errors of prior models. This is the fundamental idea behind GBT. At each iteration, the method calculates the gradient of the loss function and compares it to the prediction that the current ensemble of trees has provided. After that, it creates a new tree that is designed to imitate the negative gradient, which ultimately results in a reduction in the residual error as shown in Fig. 4.

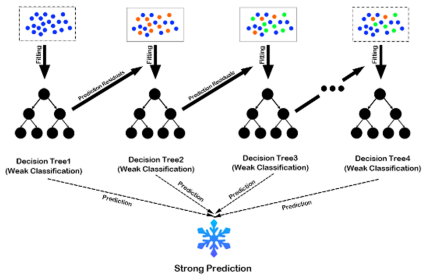



**Fig. 4.** Gradient Boosting Tree (GBT) illustration.

A classification method with an independence assumption among predictors, the Naïve Bayes classifier (NB) is based on Bayes' Theorem. A Bayes classifier, to put it simply, makes the assumption that the existence of a given feature in a class is independent of the existence of any other feature. For classification tasks like text classification, this well-liked supervised machine learning technique is employed. It simulates the distribution of inputs for a certain class or category since it is a member of the generative learning algorithm family. With this method, the algorithm can quickly and accurately produce predictions since it is predicated on the idea that the attributes of the input data are conditionally independent given the class. The new sample instance in Fig. 5 is represented by the "white" circle, and it must be assigned to the "red" or "green" class.

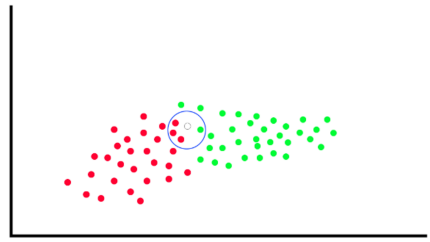


**Fig. 5.** Naïve Bayes classifier (NB) illustration.

The K-nearest Neighbors' method, or KNN for short, is made to solve regression and classification issues with ease. Its foundational principle assumes that instances that are similar to each other are situated close to each other in the feature space. KNN requires all available examples to be stored for it to work. New cases are then categorized using a similarity measure. Distance metrics like the Euclidean distribution or the Manhattan distance are frequently used to construct this similarity measure. KNN is used in classification tasks to determine which class label is more commonly used among a given data point's K-nearest neighbors. The value of K, or the number of neighbors that should be considered, is one of the most crucial factors that defines the efficacy of the model and its ability to be generalized. A larger K

value produces smoother decision borders with the possibility for a higher bias, while a smaller K number produces choice borders that may be more complex and have a higher variation. Similar to this, KNN can predict the result of a new data point in regression tasks by averaging the target values of its closest neighbors or by taking the majority vote. A simplified illustration of the K-neighbor algorithm is shown in Fig. 6. Because the ”star” obtains more ”votes” from the ”green” class, it is classified as ”red” when K = 3. However, the identical sample item is classified as ”red” with K = 6 since it now obtains more ”votes” from the ”blue” class.

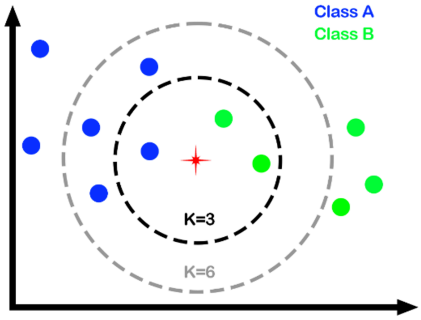



**Fig. 6.** K-nearest Neighbors (KNN) illustration.

## 2.5 Evaluations

Four key metrics—accuracy, precision, recall, and F-measure—are commonly employed to evaluate the performance of data categorization models. Two methods of data division are utilized: split testing and cross-validation. In split testing, data is divided into training and testing sets in a 70:30 ratio. Cross-validation employs the K-Fold Cross Validation method, dividing data into ten equal parts. Model performance is assessed using the confusion matrix table, as depicted in Fig. 7.

Fig. 7 displays the confusion matrix table, utilized for assessing the model's performance. A result is deemed a TP (true positive) when the forecast aligns with the actual event. False positives, or FPs, occur when accurate forecasts conflict with actual outcomes. Conversely, a TN (true negative) signifies cases where both the prediction and actual data are inaccurate. This implies that predicted outcomes do not match actual events, and in rare instances where the prediction is accurate, the event indeed occurs as predicted. In this study, TP signifies accurately forecasting graduates securing jobs. FP denotes predicting employment for a graduate who remains unemployed. TN predicts graduates remain jobless, contrary to reality. FN indicates predicting a graduate does not find employment when they have. Model performance is evaluated using the confusion matrix as follows.

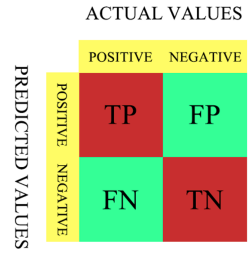



**Fig. 7.** Confusion matrix table evaluates model performance.

Accuracy is calculated as the proportion of correct predictions relative to the total predictions made, akin to Eq. (2.5).

$$Accurency = \frac{TP + TN}{TP + TN + FP + FN}. \quad (2.5)$$

Recall refers to the process of finding the percentage of data that is retrieved with all of the necessary information and satisfies the requirements. This procedure is considered to be the process of recall. in

a manner analogous to Eq. (2.6).

$$Recall = \frac{TP}{TP + FN}. \quad (2.6)$$

A technique for measuring the right number of numbers is called precision, based on the information anticipated to represent the class in question, as shown in Eq. (2.7).

$$Precision = \frac{TP}{TP + FP}. \quad (2.7)$$

As seen in Eq. (2.8), the F-Measure is a technique for averaging the Precision value, and the Recall value represents the overall performance of the response class.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \quad (2.8)$$

Specificity, also known as the True Negative Rate, is the percentage of samples that are really of the Negative Class and are projected to be of that class, as indicated in Eq. (2.9).

$$Specificity = \frac{TN}{TN + FP}, \quad (2.9)$$

$$FPR = \frac{EP}{TN + FP}, \quad (2.10)$$

$$FNR = \frac{EPFN}{TN + TP}. \quad (2.11)$$

In Eq. (2.10), the False Positive Rate (FPR) represents the fraction of samples that belong to the Negative Class but are incorrectly projected to belong to the Positive Class. According to Eq. (2.11), the False Negative Rate (FNR) is the proportion of positive class samples that are wrongly projected to be negative class samples.

The ROC AUC (Receiver Operating Characteristic Area Under the Curve) score is a widely used metric for evaluating the performance of binary classification models. It provides a single measure that summarizes the model's ability to discriminate between positive and negative classes across various threshold settings. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity), illustrating the trade-off between recall and the false positive rate. The AUC score, which ranges from 0 to 1, represents the probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative one. A model with an AUC score of 0.5 is equivalent to random guessing, while a score of 1 indicates perfect discrimination. This metric is particularly useful in imbalanced datasets where other performance measures might be misleading, making the ROC AUC score a crucial tool for assessing the robustness and effectiveness of classification models.

### 2.6 Development

Following the evaluation of the forecasting model, a graduate from Rajamangala University of Technology Lanna secured employment. The researcher intends to disseminate the findings to relevant university departments, the Office of Academic Affairs and Registrars, and admissions departments to optimize data utilization. The insights gained, including the benefits of adding majors or courses, can inform curriculum development and enhance student admissions processes.

## 3. Results and Discussions

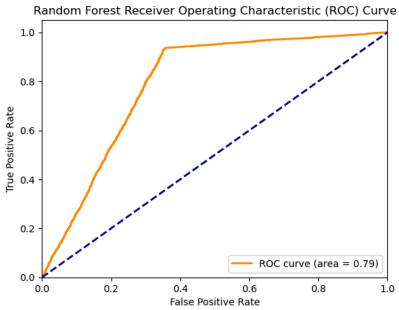
In order to conduct this study, the dataset that was utilized was collected from Rajamangala University of Technology Lanna, which is located in Chiangmai, Thailand. A total of 4,352 samples of data that are important to the employability and academic performance of students during the academic year 2023 are included in the dataset. Following the con-

clusion of split testing, the test results are presented sequentially for clarity. In addition, it is essential to compare the results across the board when the process of projecting the data and performance is being carried out. To ensure meaningful and reliable interpretation, multiple performance metrics—including accuracy, recall, precision, F1-score, and AUC—were jointly analyzed rather than relying on a single indicator.

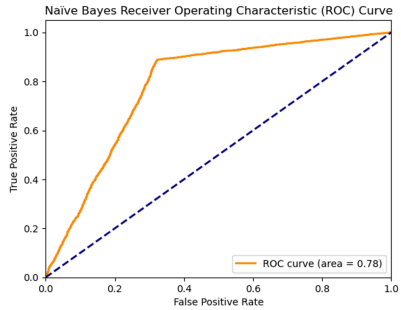
At the center of the performance analysis of any model that is derived from the ROC curve is the value of the AUC, which stands for the area under the curve. We consider this number as a key signal to determine the model's ability to differentiate between distinct classes. The fact that the area under the curve (AUC) value has increased indicates that the model is more successful in distinguishing between positive and negative classes, as demonstrated in Figs. 8(a) – 8(e). However, AUC alone cannot fully represent model stability in real-world prediction tasks; therefore, the trade-off between precision and recall must also be considered. In Fig. 8(a), Random Forest has 0.784 AUC. Random Forest has the second-highest AUC, indicating strong class differentiation. Random Forest has the best F1-score, indicating a balance between accuracy and recall, but its AUC is lower than Gradient Boosted Trees'. This suggests that RF generates more stable predictions due to its ensemble of decorrelated trees, which reduces overfitting and handles heterogeneous educational data more effectively than GBT.

In Fig. 8(b), the Naive Bayes method has an AUC of 0.783. AUC values are higher than decision trees and equivalent to K-nearest Neighbors. The Naive Bayes model can distinguish classes; however, it is not the most accurate. The decision trees in Fig. 8(c) have an AUC of 0.778. Its lowest AUC indicates that it can differentiate lower classes relative to other models. However, decision tree's accuracy and F1-score values remain acceptably distinct, indicating that the model is still successful in some cases. In Fig. 8(d), the K-nearest Neighbors has an AUC of 0.783. The AUC value matches the Naive Bayes method, indicating excellent class differentiation. Despite a lower AUC value, K-Nearest Neighbors has a higher F1-score than Naive Bayes. We get 0.813 AUC for gradient-boosted trees. Its area under the curve (AUC) value is the highest of all models, indicating that it can distinguish classes well. As shown in Fig. 8(e), a high area under the curve (AUC) value suggests that true positive and true negative prediction rates are higher than false positive and false negative prediction rates. However, the sequential boosting process makes GBT more sensitive to noisy or imbalanced datasets, explaining why GBT—despite having the highest AUC—still yields a lower F1-score and less consistent performance than RF.

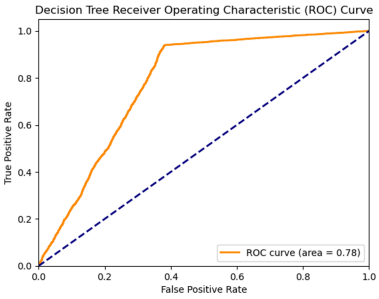
AUC measurements of the ROC curve allow us to draw inferences. Gradient-boosted trees have a high AUC and stand out as a model that can distinguish the best-performing classes among all models. Random Forest has the highest F1 score and AUC. This makes it a useful memory aid for balancing accuracy and recall. Naive Bayes and K-Nearest Neighbors have equal AUCs, indicating they can distinguish classes well. Decision trees have the lowest AUC but high accuracy and F1-scores, making them a favorable choice in some cases. If you need a model that can distinguish high classes, gradient-boosted trees may be best. However, if you want a model that balances precision and recall, a random forest may be better.



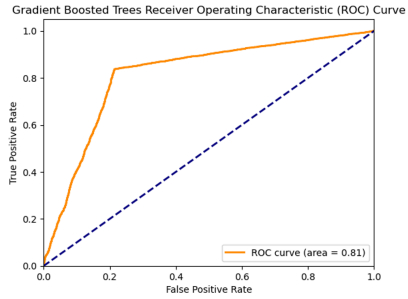

(a). Random forest model ROC curve




(b). Naive Bayes model ROC curve




(c). Decision tree model ROC curve




(d). K-Nearest Neighbors model ROC curve

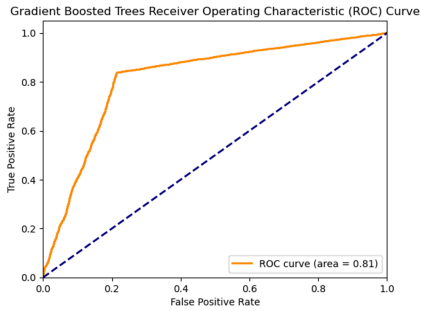



**Fig. 8.** Five models with a ROC curve.

Taken together, these results suggest that GBT excels in ranking capability, while RF provides the most consistent classification performance. For practical employability prediction—where both misclassification types matter—RF's balanced precision–recall behavior makes it a more reliable model than GBT despite its lower AUC.

Five distinct machine learning models are compared in Table 2 based on performance metrics: Random Forest (RF), Gradient Boosting Trees (GBT), Naive Bayes (NB), K-Nearest Neighbors (KNN), and Decision Trees (DT). Accuracy, recall, precision, F1-score, and area under the curve (AUC) are the measures that are taken into consideration.

**Table 2.** Consider each model's overall performance.

| Model | Accuracy | Recall | Precision | F1-score | AUC |
|---|---|---|---|---|---|
| RF | **83.72** | 78.98 | **83.68** | **81.26** | 0.784 |
| GBT | 81.96 | **81.11** | 79.75 | 80.42 | **0.813** |
| NB | 81.66 | 78.21 | 79.92 | 79.06 | 0.783 |
| KNN | 82.06 | 78.03 | 80.77 | 79.38 | 0.783 |
| DT | 83.12 | 77.89 | 83.42 | 80.56 | 0.778 |

Random Forest (RF): With a good

F1-score of 81.26, RF achieves the highest accuracy (83.72%) and precision (83.68%). Its recall rate is 78.98%, and its AUC is a reasonable 0.784. Gradient Boosting Trees (GBT) show the greatest AUC (0.813) and strong recall (81.11%). It has an F1-score of 80.42, accuracy of 81.96%, and precision of 79.75%. With an accuracy of 81.66%, recall of 78.21%, precision of 79.92%, and an F1-score of 79.06, Naive Bayes (NB) exhibits impressive results. It has an AUC of 0.783. K-Nearest Neighbors (KNN) has an F1-score of 79.38, accuracy of 82.06%, recall of 78.03%, and precision of 80.77%. It has an AUC of 0.783. Decision Trees (DT) has the lowest AUC of 0.778 out of all the models, with an accuracy of 83.12%, precision of 83.42%, F1-score of 80.56, and recall of 77.89%.

Overall, both DT and RF demonstrate higher levels of accuracy and precision, with RF having a slightly higher F1-score than DT. It is possible that GBT is a strong contender because to its greater recall and AUC; however, this will rely on the usage and the significance of each metric. Based on a holistic evaluation of all performance indicators, the Random Forest model is the most suitable algorithm for employability prediction, as it provides strong accuracy, precision, and F1-score while demonstrating greater robustness to overfitting compared to GBT.

## 4. Conclusion, Limitations and Future Research Ideas

This study compared five machine learning algorithms—Decision Tree, Random Forest, Gradient Boosted Trees, Naïve Bayes, and K-Nearest Neighbors—to predict the employability of graduates from Rajamangala University of Technology Lanna (RMUTL). Among the models, Random Forest demonstrated the best overall performance, particularly in accuracy and F1-score, making it the most effective approach for practical employability prediction. Although Gradient Boosted Trees achieved the highest AUC, its lower stability across metrics suggests that Random Forest offers a more balanced and reliable solution.

The findings are consistent with prior research in educational data mining. Kotsiantis [17] reported the usefulness of Decision Tree and Naïve Bayes for student outcome prediction. Natek [18] highlighted the strength of Random Forest in modeling complex educational datasets, while [19] and [20] emphasized the growing importance of data mining and ensemble learning for supporting academic decision-making. The alignment between the present results and previous evidence reinforces the suitability of ensemble-based models—particularly Random Forest and Gradient Boosted Trees—for employment prediction in higher education contexts.

### 4.1 Limitations

There are some limitations to this study, although the results are promising. First, the study was conducted within a single institution, and therefore the findings may not be generalizable to other universities or regions. A second limitation is that the study did not explicitly address the potential issue of class imbalance in employment status, which may have influenced the accuracy and comparative performance of the models. In addition, the study does not provide detailed information about the parameter tuning process for each model, which may raise questions regarding the fairness and consistency of the comparison.

### 4.2 Future Research Ideas

1) Future studies may expand the dataset to include additional academic years in order to improve the generalizability of the model.

2) explored, as these factors may further contribute to employability outcomes.

3) Methods to address class imbalance, including SMOTE and cost-sensitive learning, should be applied to enhance the stability and fairness of model performance.

4) A job or career guidance application based on the best-performing model could be developed and made available for use by career services or academic advisors.

## 5. Acknowledgements

The authors would like to express their appreciation to all individuals who contributed to this research. Special thanks are extended to Rajamangala University of Technology Lanna for providing the essential datasets used in this study. The authors also gratefully acknowledge the valuable guidance, constructive feedback, and continuous support from advisors and collaborators throughout the research process.

## References

[1] Hehn TM, Kooij JF, Hamprecht FA. End-to-end learning of decision trees and forests. Int J Comput Vis. 2020;128(4):997-1011.

[2] Aria M, Cuccurullo C, Gnasso A. A comparison among interpretative proposals for Random Forests. Mach Learn Appl. 2021;6:100094.

[3] Vu QV, Truong VH, Thai HT. Machine learning-based prediction of CFST columns using gradient tree boosting algorithm. Compos Struct. 2021;259:113505.

[4] Itoo F, Meenakshi M, Singh S. Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. Int J Inf Technol. 2021;13(4):1503-11.

[5] Dong Y, Ma X, Fu T. Electrical load forecasting: A deep learning approach based on K-nearest neighbors. Appl Soft Comput. 2021;99:106900.

[6] Dwivedi YK, Hughes L, Ismagilova E, Aarts G, Coombs C, Crick T, et al. Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. Int J Inf Manag. 2021;57:101994.

[7] Hassani H, Silva ES, Unger S, TajMazinani M, Mac Feely S. Artificial intelligence (AI) or intelligence augmentation (IA): what is the future? AI. 2020;1(2):8.

[8] Zhang Z, Zhu X, Liu D. Model of gradient boosting random forest prediction. In: Proceedings of the IEEE International Conference on Networking, Sensing and Control (ICNSC); 2022. p. 1-6.

[9] Dong M, Yao L, Wang X, Benatallah B, Zhang S, Sheng QZ. Gradient boosted neural decision forest. IEEE Trans Serv Comput. 2021;16(1):330-42.

[10] Ayyadevara VK. Gradient boosting machine. In: Pro machine learning algorithms. Berkeley: Apress; 2018. p. 465-506.

[11] Roy SS, Chopra R, Lee KC, Spampinato C, Mohammadi-Ivatloo B. Random forest, gradient boosted machines and deep neural network for stock price forecasting: a comparative analysis on South Korean companies. Int J Ad Hoc Ubiquitous Comput. 2020;33(1):62-71.

[12] El Boujnouni M. A study and identification of COVID-19 viruses using N-grams with Naïve Bayes, K-nearest neighbors, artificial neural networks, decision tree and support vector machine. In: Proceedings of the International Conference on

Intelligent Systems and Computer Vision (ISCV); 2022. p. 1-7.

[13] Schröer C, Kruse F, Gómez JM. A systematic literature review on applying CRISP-DM process model. Procedia Comput Sci. 2021;181:526-34.

[14] Rolke W, Gongora CG. A chi-square goodness-of-fit test for continuous distributions against a known alternative. Comput Stat. 2021;36(3):1885-1900.

[15] De Sousa MS, Veiga CE, Albuquerque RD, Giozza WF. Information gain applied to reduce model-building time in decision-tree-based intrusion detection system. In: Proceedings of the Iberian Conference on Information Systems and Technologies (CISTI); 2022. p. 1-6.

[16] Aloisio M, Angeletti P, Casini E, Colzi E, D'Addio S, Oliva-Balague R. Accurate characterization of TWTA distortion in multicarrier operation by means of a correlation-based method. IEEE Trans Electron Devices. 2009;56(5):951-8.

[17] Kotsiantis S, Pierrakeas C, Pintelas P. Predicting student performance in distance learning using machine learning techniques. Appl Artif Intell. 2004;18(5):411-26.

[18] Natek S, Zwilling M. Student data mining solution–knowledge management system related to higher education institutions. Expert Syst Appl. 2014;41(14):6400-7.

[19] Baradwaj BK, Pal S. Mining educational data to analyze students' performance. arXiv. 2012; arXiv:1201.3417.

[20] Romero C, Ventura S. Educational data mining: a review of the state of the art. IEEE Trans Syst Man Cybern C Appl Rev. 2010;40(6):601-18.