International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Predicting the Employability of Filipino Students Using Mock Job Interview Data: A Comparative WEKA Classifier Analysis

Pardo Alven P[1], Mosa Jhedelson A[2], Barol Rhebie D[3], Catuburan Shane Ray B[4]

[1,2,3,4]Student Researcher, Bachelor of Science in Information Systems,Surigao Del Norte State University

Abstract

The research investigates how data mining methods can predict whether Filipino students will secure jobs through testing these methods with the WEKA machine learning platform. The research evaluated four algorithms which included Naïve Bayes, SMO, IBk, and Logistic Regression through 10-fold cross-validation using 2,982 mock job interview records obtained from university agencies throughout the Philippines. The results show that SMO is the highest-performing classifier, achieving 92.54% accuracy, followed by IBk at 89.97%, Logistic Regression at 88.45%, and Naïve Bayes at 85.21%. The most important soft-skill attributes which interviewers based their assessments on included Mental Alertness and Communication Skills and Self-Confidence and Ability to Present Ideas, but the dataset showed Communication Skills and Mental Alertness as the most effective attributes for class separation. SMO showed the best performance through its predictive capabilities, but the system lacked explainability, whereas IBk established dependable performance, and Logistic Regression delivered an easily understandable model with similar accuracy results, while Naïve Bayes operated as the weakest basic model due to its easy-to-use nature. Davao Region higher education institutions should use SMO-based predictive modeling along with employability scoring and targeted soft-skills interventions to identify at-risk students through their early identification process which will enhance mock interview training and career guidance and pre-employment preparation.

Keywords: Filipino Students, Employability Prediction, WEKA, Data Mining, Machine Learning, Mock Job Interview

1. Introduction

Data mining and machine learning advancements have transformed various fields, including the education industry. The analytical frameworks of EDM enable organizations to predict students' results while detecting students who require educational assist and provide base data for career and academic coaching programs, which use databased approaches to support students [4]. The Philippine context shows that recent modeling on employability using mock interview data for machine learning identification of student attributes for successful workforce readiness [3].

Data mining techniques including Naïve Bayes, SMO, IBk, and Logistic Regression are some of the most

widely used in employability analysis. These algorithms can analyze the massive volume of mock interview data and find the latent correlations between soft skills and employability status of Filipino students. However, a single best fit model remains to be a concern as several techniques and approaches have been previously employed with varying degrees of assessing accuracy and interpretingability in data mining solutions.

The current study is to compare the many data mining algorithms using WEKA on the Students' Employability Dataset – Philippines. Based on the mock job interviews records conducted, with data collected from different university agencies in the Philippines (particularly in Davao Region), this paper aims to determine the best classification model concerning Students' Employability Dataset – Philippines. Besides, the impact of key soft-skill factors in predicting positive or negative employment outcomes is also investigated locally.

The findings of this study will aid the educational institutions of Davao City and the Davao Region in formulating data-driven strategies to capacitate students at risk of low employability.

2 Problem Statement

In this regard, employability predictions for students based on various factors remain the most pertinent challenge in the Philippine educational sector, with soft skills being one of the critical factors and mock job interviews serving as a tool for assessment (CHED, 2025). Traditional methods of assessment do not give much insight into employability and are based on manual subjective assessment with no prediction. In the absence of a predictive model, universities in the Davao Region cannot ascertain students who may exhibit characteristics of 'Less Employable,' thereby limiting the efficacy of intervention strategies.

The case of employability essays and decision in education is a chance to improve it through the application of data mining techniques. However, identifying the best WEKA classifier is difficult, as there are high levels of accuracy and interpretability with some classifiers over others. To fill the gap, this study will compare various data mining techniques based on WEKA classifiers in predicting the employability of Filipino students.

Understanding the key predictors of employability is essential for higher education institutions in the Philippines in proactively developing strategies that improve the outcome of graduates. This study will harness machine learning to gather relevant insights from the raw mock interview data and hopefully contribute to the improvement of data-driven career support systems for higher education institutions in Davao City and the Davao Region.

3 Objectives

To start with, the overall objective of this study is to compare different data mining techniques using WEKA in order to choose the best algorithm to conduct analysis with on data based on mock job interview and employability of Filipino students.

The specific objectives are:

1. To evaluate the predictive accurateness of different data mining algorithms such as Naïve bayes, SMO and IBk and logistic regression in predicting employability.
2. To distinguish the most significant mock interview attributes that influence mental alertness, communication skills, and self-confidence with employability.
3. To compare the effectiveness of the classification models based on the prediction of the employability status, using accuracy, kappa statistics, precision, recall, F-measure and ROC area.

4. To assess the interpretability and usability of different data mining techniques in the Philippine higher education context, so as to provide insights to educators and policymakers.
5. To come up with recommendations on the implementation of data-driven strategies of universities in Davao Region to improve the mock interview training and employability.

The study will answer the following research questions.

1. Which data mining algorithm yields the highest accuracy in predicting the employability of Filipino students?
2. What are the major attributes of employability skills?
3. How do the different WEKA classifiers perform and interpret?
4. What data-driven recommendations for at-risk students in Davao Region?

4 Scope of the Study

This study is on the basis of a data mining classification technique application on the machine learning WEKA tool for predicting the employability of Filipino students, based on a mock job interview. These are the boundaries for the scope:

1. Dataset Coverage. For the study, the Students' Employability Dataset – Philippines with a total of 2,982 student records is used. The attributes contained in the dataset include general appearance, manner of speaking, physical condition, mental alertness, self-confidence, ability to present ideas, communication skills, and student performance rating, with employability status as the target class.
2. Algorithm Scope. First is the scope of the algorithm in the study, with the four chosen WEKA classifiers selected to classify whether a student is Employable or LessEmployable.
3. Evaluation Metrics. The effectiveness of the algorithms is measured using the standard classification measures of correctly classified instances, incorrectly classified instances, kappa statistics, precision, recall, F-measure, ROC area, meaning absolute error, and root mean squared error.
4. Temporal Scope. Four: temporal scopeThe data analyzed in this study is historical mock job interview data and does not include real-time monitoring or actual post-graduation employment outcomes.
5. Contextual Scope. The scope contextualizes limitation in terms of employability predication in the Philippine higher educational context using records gathered by university agencies in the Philippines, with their relevance being emphasized in the Davao Region institutions.

Limitations of the Study

However, there are certain limitations associated with this study.

1. Limitation of Datasets. Other than that dataset contains 2,982 records, but limited to mock job interview evaluations, which may not reflect actual hiring decisions or labor-market outcomes.
2. Feature Limitation. Therefore, the attributes for analysis are limited to variables based on an interview. Demographic background, academic achievement, socioeconomic condition, and technical examination results are some of the possible predictors that are not included.
3. Algorithm Limitation. The study is limited to the comparison between Naïve Bayes, SMO, IBk, and Logistic Regression. No other machine learning algorithms were analyzed.
4. Generalizability Limitation. Therefore, the dataset is limited to the context of the Philippines and cannot be generalized to other geographical contexts, institutions, or countries.
5. Outcome Limitation. The findings have been based on predicted employability from the mock interview performance only and should not be interpreted as a direct measure of actual job placement

or long-term career success.

5 Related Work

The integration of educational data mining (EDM) to forecast workforce readiness has earned considerable interest. All related literature reviewed in this section was strictly published within the last five years (2020 to present) to ensure such predictive models align with industry standards presently in place post-pandemic.

Machine Learning in Employability Prediction (2020-Present)

Several recent studies have validated the effectiveness of machine learning to predict student employability. Casuat and Festijo [1] demonstrated the efficacy of classification techniques in forecasting engineering students' employability noting that decision trees and ensemble methods tend to outperform statistical approaches in classifying intricate student attributes. In the same vein, Moumen et al. [2] over a comparison of machine learning algorithms concluded that forecasting allows educational administrators to monitor the variables that correlate with workforce readiness in populations, facilitating targeted interventions in the early stages for at-risk students.

Contextualizing the Philippine Setting

In the same of the Philippines, there is now a growing demand to link higher education outcomes to the actual needs of its labor market to address the issue of underemployment. Using a similar Philippine mock interview dataset, Olipas [3]. This is in line with a global shift identified by Saidani et al. [5], who used deep learning and big data analytics on student internship data to show that employing context-aware predictive models can significantly improve institutional career counseling.

Gaps in Current Literature

But despite the increasing use of machine learning to predict employability, there are several gaps;

1. Limited classifier comparison - Many studies focus on a single algorithm, with fewer studies comparing Naïve Bayes, SMO, IBk, and Logistic Regression classifiers altogether in one study.
2. Limited emphasis on interview-based soft skills – There are more emphasis on academic or demographic factors in most studies, with limited focus on attributes like Mental Alertness, Self-Confidence, Ability to Present Idea, and Communication Skills.
3. Limited Philippine-based studies - There is still a lack of localized studies with a Philippine dataset, especially in the field of student employability prediction.
4. Need for practical application - Practically, many studies showcase accuracy, but few demonstrate the application of these models in schools to guide career paths and intervene with students.

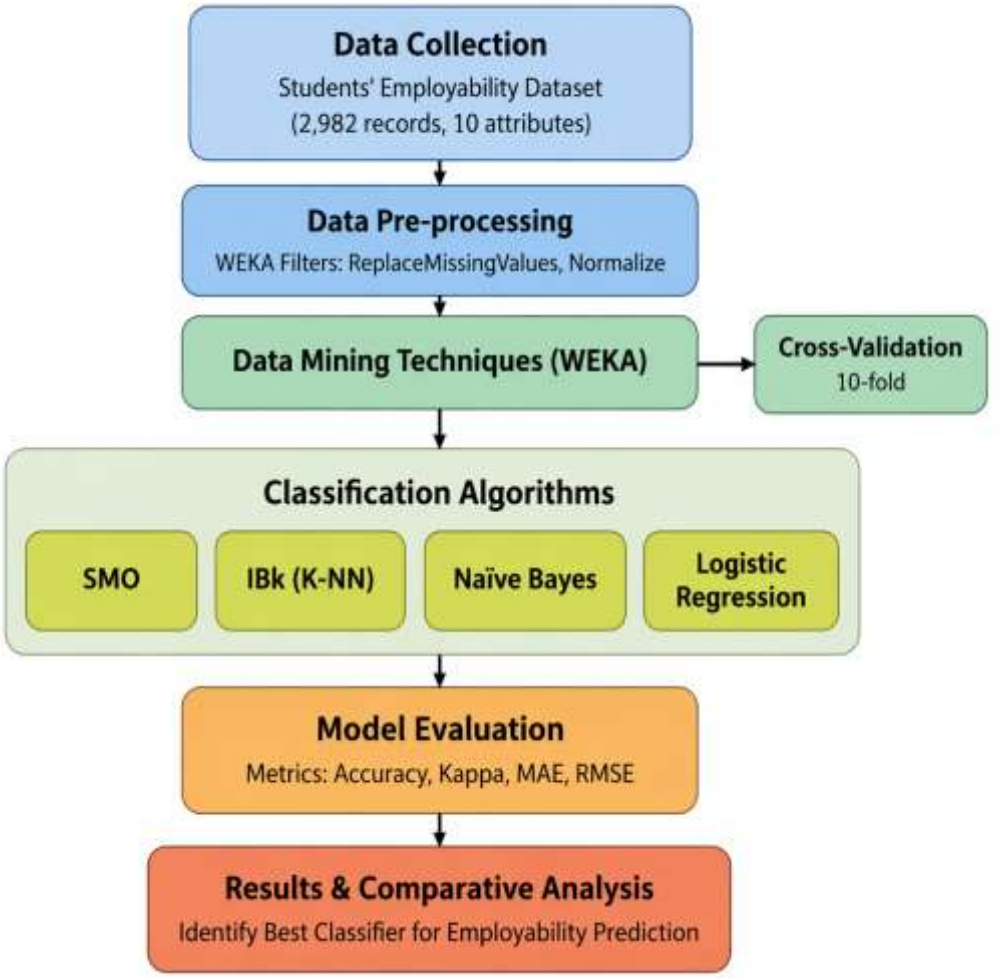
Addressing these gaps will enhance the empirical evidence available for data mining applications in predicting employability, assisting higher education institutions in adopting effective data-driven approaches to boost their students' readiness for the workforce.

6 Methodology

This section shows the method to be used to analyze and predict student employability by applying different machine learning classifiers. It allows us to identify recurring patterns, correlations, and factors that affect workforce readiness. Employing various algorithms like SMO (Support Vector Machine), IBk (K-Nearest Neighbors), Naïve Bayes, and Logistic Regression facilitates comparison, guaranteeing the selection of the most accurate model for classification and prediction. The datasets are tested using cross-

validation, and output evaluations are conducted from the classifiers to obtain meaningful and reliable results.

Figure 1: Data Mining Process Flow of the Study



6.1 Data Collection

Sources of Data

The study used the "Student-Employability-Datasets"—a publicly accessible repository containing anonymized assessments of university students during institutional mock job interviews in the Philippines.

Data Collection Methods and Tools

There are 2,982 instances (student records) and 10 attributes in the dataset – Name of Student, General Appearance, Manner of Speaking, Physical Condition, Mental Alertness, Self-Confidence, Ability to Present Ideas, Communication Skills, Student Performance Rating, and CLASS, the target variable (Employable vs LessEmployable). The dataset was consolidated into a CSV/ARFF format for WEKA compatibility.

6.2 Data Pre-processin

Data Cleaning and Preparation

To make sure the dataset remains accurate and consistent, it was cleaned up properly before the analysis

process began. Other irrelevant attributes like "Name of Student" were also removed because they are not useful in predictive modeling. Missing values have been checked, and there are no significant null values in the dataset. Inconsistent labels (such as Employable and Less Employable) were ensured to be uniform. Categorical attributes such as General Appearance, Communication Skills, and Mental Alertness were converted into numerical or nominal form compatible with WEKA. Lastly, to ensure smooth data processing within the tool, the dataset was saved in ARFF format.

In addition, attribute selection was done to ensure only relevant attributes on employability are considered, thus eliminating redundant attributes and further boosting the performance of the model.

Handling Outliers

Potential outliers that will distort the efficiency of machine learning algorithms were identified and dealt with in the dataset. To examine the Student Performance Ratings further, this study performed statistical analysis and found extreme high or low values in some affecting numerical attributes. The z-score method was used to check for unusually high or low values, with the threshold set at ±3.

Also, the boxplot visualization in WEKA was used to check for the distribution of attributes and provided insight into skewness that might be present in mock interview scores collected from universities within the Philippines, more specifically in the Davao Region.

To reduce the impact of extreme values on model performance a Winsorization approach was employed by replacing the extreme data with the closest legitimate values within an acceptable range. The classification models to be trained using the dataset include Naïve Bayes, SMO (Support Vector Machine), IBk (Instance-Based K-nearest Neighbor), and Logistic.

Data Normalization

To ensure that all numerical attributes contributed equally to the classification process, min-max normalization was conducted on the dataset. The technique transforms values to fall within a uniform range of between 0 and 1 thus making sure that attributes with more numerals do not dominate the model. The formula for normalization that was used in this study is given below;

Equation 1: $x' = (x - \min(x)) / (\max(x) - \min(x))$

Normalization was particularly applied on the relevant numerical attributes such as for the performance ratings of the students to bring consistency in the entire data set. This is seminal in enhancing the performance of distance and margin-based algorithms; IBk and SMO, respectively.

6.3 Data Mining Techniques

Data mining techniques are crucial for identifying patterns from mock-job interview results. For classification, predictive models were developed to predict student employability status from attributes related to mock job interviews. Classification is appropriate because they cluster student records under Employable and LessEmployable classes, so that institutions can identify students who may need to be additionally intervened with, to boost soft-skills relevant to job interviews.

Classification Techniques

1. The Naive Bayes is a probabilistic classifier based on Bayes theorem and assumes the independence of attributes. Therefore, Naïve Bayes classifier was implemented and evaluated on the dataset alongside other classification algorithms.
2. Sequential Minimal Optimization or SMO is the WEKA implementation of Support Vector Machine. The classes are separated by maximizing the margin between Employable and LessEmployable cases through an optimal hyperplane. This classification algorithm was chiefly selected based on its performance in high dimensionality classification problems.

3. IBk is the nearest neighbor or instance-based learner that classifies a case based on the class of its nearest training instances. It is useful in this study because it can identify employability status based on similarity among student interview profiles.
4. Logistic Regression is a statistical classification model for dichotomous variables. It gives the probability that the student belongs to either the Employable or the LessEmployable class, hence is appropriate for interpreting the employability prediction in terms of probability.

Table 1: Data Mining Techniques Used

Algorithm	Description	Purpose in Study
Naïve Bayes	Probabilistic classifier based on Bayes' theorem	Used as baseline model for comparison
SMO (Support Vector Machine)	Separates classes using optimal hyperplane	Main model for high accuracy prediction
IBk (K-Nearest Neighbors)	Classifies based on closest training instances	Evaluates similarity-based prediction
Logistic Regression	Statistical model for binary classification	Measures probability of employability

Table 2: Comparison of Classification Techniques Used in the Study

Algorithm	Accuracy	Interpretability	Handling of Missing Data	Computational Efficiency
Naïve Bayes	Moderate	Moderate	High	Very Fast
SMO (Support Vector Machine)	Very High	Low	Low	Moderate
IBk (K-Nearest Neighbors)	High	Moderate	Moderate	Moderate
Logistic Regression	High	High	Low	Moderate

Naïve Bayes, SMO, IBk, and Logistic Regression are purposely selected for a balanced comparison of probabilistic, margin-based, distance-based, and statistical classification approaches. Using these techniques, the study will identify what model best predicts employability from mock interview data.

6.4 Tools and Technologies Used

For the implementation of this study, there were appropriate tools to analyze and interpret the student's employability data, which was based on mock job interviews held in the Philippines, specifically in the Davao Region.

The primary software used in this research is WEKA (Waikato Environment for Knowledge Analysis), which is a machine learning open-source tool developed by the University of Waikato. For this, a GUI has been provided for performing data mining tasks such as classification, clustering, and regression [7].

WEKA consists of preprocessing tools for data cleaning, transformation, and normalization, with which the dataset was prepared. Bayes, SMO, IBk, and Logistic Regression classification algorithms that were

used in this study. Besides, the platform provides accuracy, kappa, precision, recall, and confusion matrix as evaluation metrics to allow the performance to be compared effectively using 10-fold cross-validation.

7 Data Analysis

The dataset was analyzed in WEKA using Naïve Bayes, SMO, IBk, and Logistic Regression, with 10-fold cross-validation to ensure consistent evaluation across models. The analysis showed a stark difference in the quality of prediction. SMO achieved the best accuracy results and kappa values, followed by IBk. Logistic Regression produced moderate results, while Naïve Bayes registered the lowest overall accuracy out of the selected classifiers.

The confusion matrices also show that the Employable class was more consistently classified than the LessEmployable class. Although, some of the interview-related attributes cover both categories and separating them by such information makes the LessEmployable group more difficult to separate. The data analysis, in general, indicates that margin- and distance-based classifiers were better than the probabilistic baseline for this particular data set. Numerical comparison in detail is given in the Results section under Different Algorithm Classification, as the case stands following the flow of the sample research template.

As shown in the confusion matrices, the models consistently classified the Employable class, whereas prediction of the LessEmployable class was problematic. Some interview-related attributes were observed to overlap in both categories; hence this pattern was expected to confuse the classifier. To compare the models systematically, Table 3 summarizes the main performance metrics used to evaluate each classifier.

Key Insights from the Data

SMO scored the highest classification accuracy of 92.54%, thus making it the strongest model for employability prediction in this study. IBk was not left behind, recording an accuracy rate of 89.97% and a kappa statistic value of 0.7927, which implies strong agreement between the predicted and actual classes. It had the lowest accuracy of 85.21% and it was the least suitable classifier for this study because it assumes the independence of attributes, whereas interview variables could be naturally related to one another, for example, communication skills and manner of speaking.

8 Results

Dataset of Employability Ratings

The dataset includes 2,982 mock job interviews, with eight ordinal employability attributes and one binary target class. The target class for student employability consists of 1729 students being Employable (57.98%) and 1253 students labeled LessEmployable (42.02%). The ordinal employability attributes include General Appearance, Manner of Speaking, Physical Condition, Mental Alertness, Self- onfidence, Ability to Present Ideas, Communication Skills, and Student Performance Rating. This distribution ensures a data set that can be rubbed for patterns and comparison of classifiers.

Figure 2: Interpretation of Mental Alertness (MA) vs. Employability

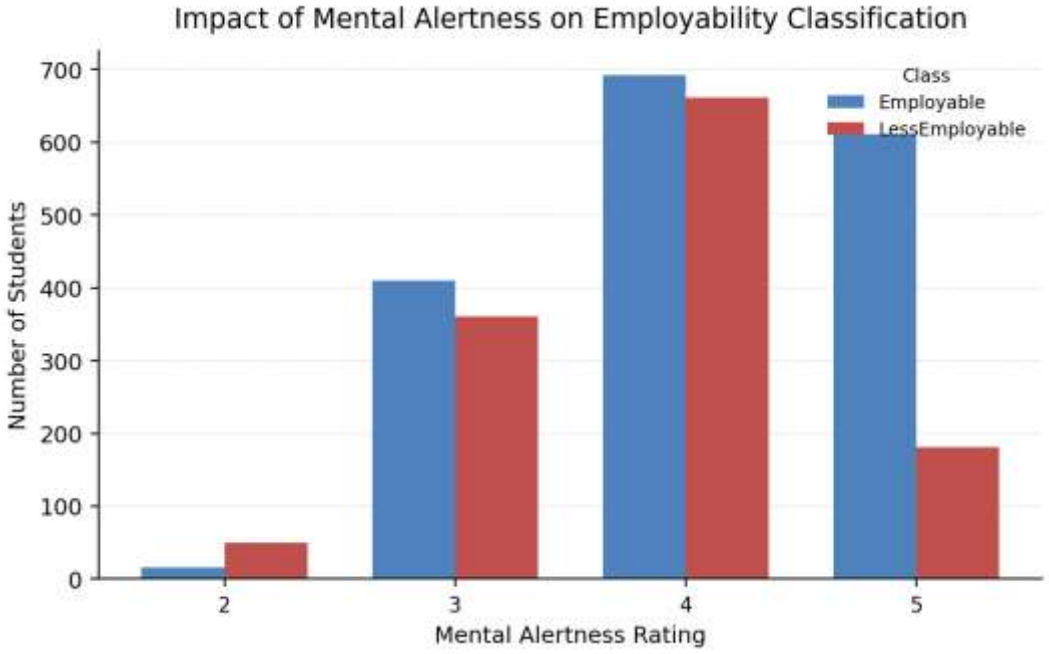


Figure 2: Distribution of Employability Classes across Mental Alertness Ratings The blue bars represent Employable students against the LessEmployable ones represented by red bars. The pattern shows that rating 5 is strongly dominated by Employable students, whereas rating 2 is dominated by LessEmployable students. Employability bars remain more mixed, though the blue still tends to exceed red. In overall, mental alertness is positively associated with employability and is a strong discriminator between the two classes.

Figure 3: Interpretation of Self-Confidence (SC) vs. Employability

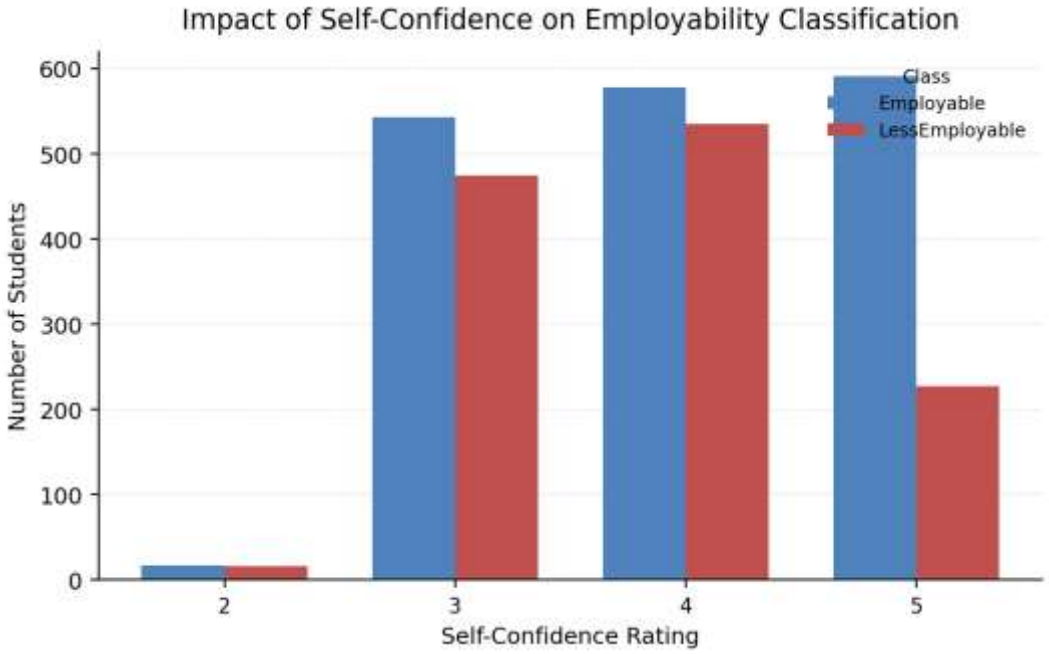


Figure 3 presents the relationship between Self-Confidence and employability classification. The colors remain to represent employability classification, with blue for Employable and red for LessEmployable.

Given that each rating value is an integer, the bars show that the maximum gap is at 5, with Employable students outnumbering LessEmployable students. At mid-range ratings a pattern is more balanced, which suggests that Self-confidence contributes to employability but becomes most meaningful at the highest level. The graph therefore treats self-confidence as a positive indicator of workforce readiness.

Figure 4: Interpretation of Ability to Present Ideas (API) vs. Employability

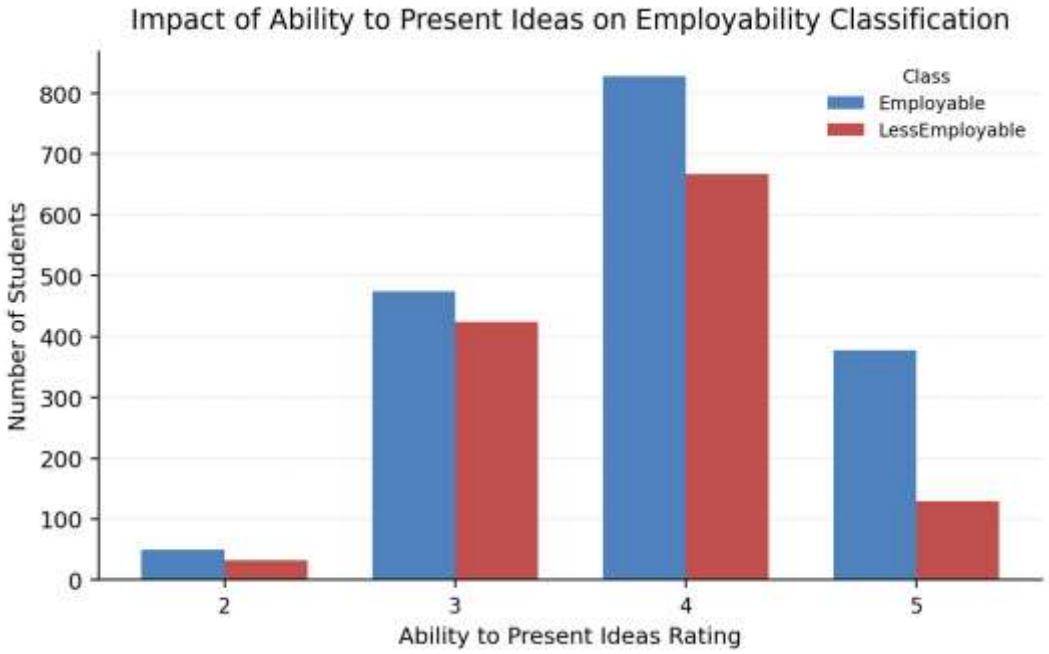


Figure 4 shows how the Ability to Present Ideas changes in different classes of employability. Blue bars represent the Employable cases, whereas the red bars represent the Less Employable cases. It is noted that larger numbers of Employable students correspond to rating 5 for demonstrating ideas. Ratings 3 & 4 remain mixed, but the employs preferably remain high at these levels too. This ability to present ideas is quite a meaningful factor over the prediction on employability.

Figure 5: Interpretation of Communication Skills (CS) vs. Employability

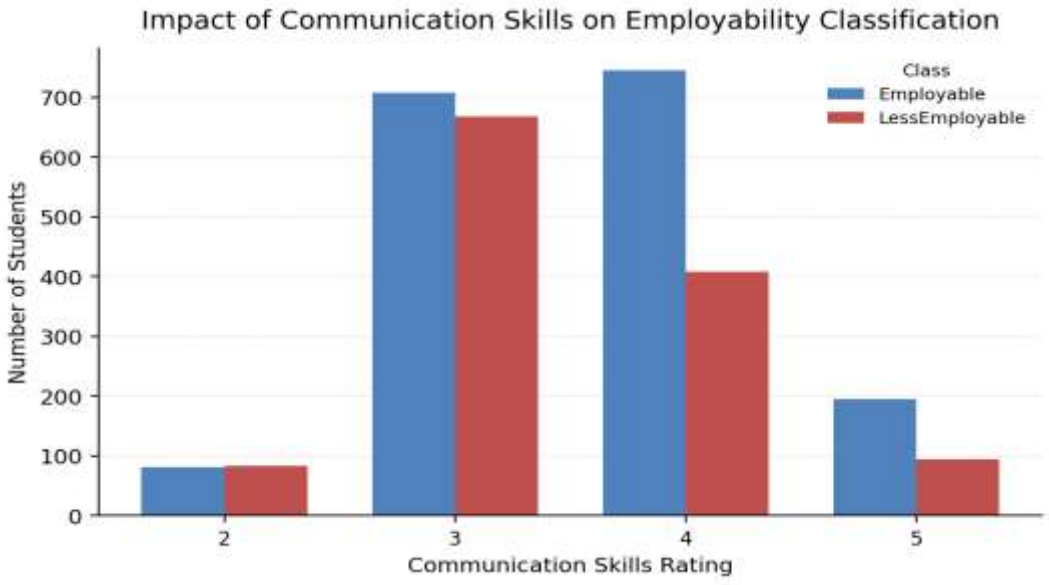


Figure 5 shows the distribution of Employable and LessEmployable students by Communication Skills rating. This graph presents the clearest case separation pattern out of the whole data set. The bars indicating employable students (blue) are much taller than the bars showing less employable students (red) at these levels of ratings, suggesting that strong communication skills make a student employable. By contrast, 2 is almost equal but with a little dominance by the LessEmployable class. The diagram can, therefore, be said to prove that communication skills are the most influential indications of Students' employability statuses.

Different Algorithm Classification

In this study, four classification algorithms- Naïve Bayes, SMO, IBk, and Logistic Regression were assessed under WEKA for student employability analysis. The models were tested using the same dataset and validation setting for a fair comparison. A comparison matrix provides a summary of the major performance matrix of the classifiers in Table 3 with the succeeding WEKA visualization showing how each classification algorithm separated the Employable and LessEmployable classes.

Table 3: Comparison of Classification Models in Predicting Employability

Algorithm	Correctly Classified (%)	Incorrectly Classified (%)	Kappa Statistic	General Interpretation
SMO	92.54%	7.46%	0.8412	Highest-performing model
IBk (K=1)	89.97%	10.02%	0.7927	Strong and reliable classifier
Logistic Regression	88.45%	11.55%	0.7610	Moderately effective classifier
Naïve Bayes	85.21%	14.79%	0.6985	Baseline model with lowest accuracy

In addition, Figures 6 to 9 show the WEKA visualizations of the selected algorithms. The figure-based results further prove the numerical comparison in Table 3.

Figure 6: Margin Curve Visualization of SMO

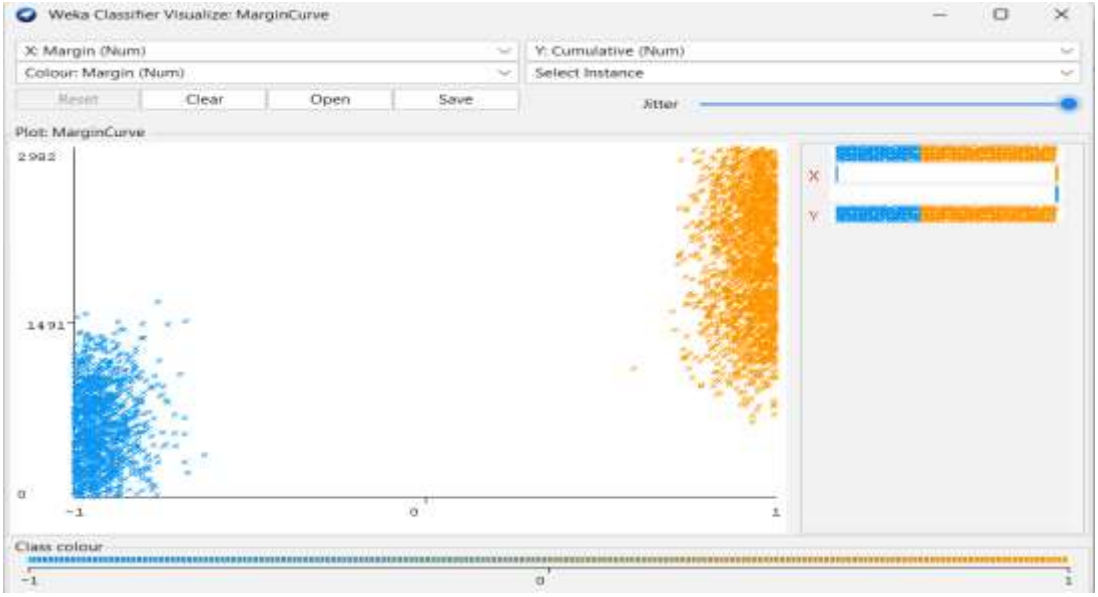


Figure 6 is the margin curve produced in WEKA for SMO. The x-axis shows the margin value, while the y-axis shows the sum of instances. The gradient starts from blue close to -1 and orange close to +1. Negative margins, indicative of incorrect classifications, were shown by blue gradient values on the left side, while positive margins, indicative of accurate classifications, were shown by orange gradients on the right side. The concentration of instances towards positive margin indicates that the SMO classifier produced more correct predictions than uncertain and incorrect predictions.

Figure 7: IBk Classifier Results

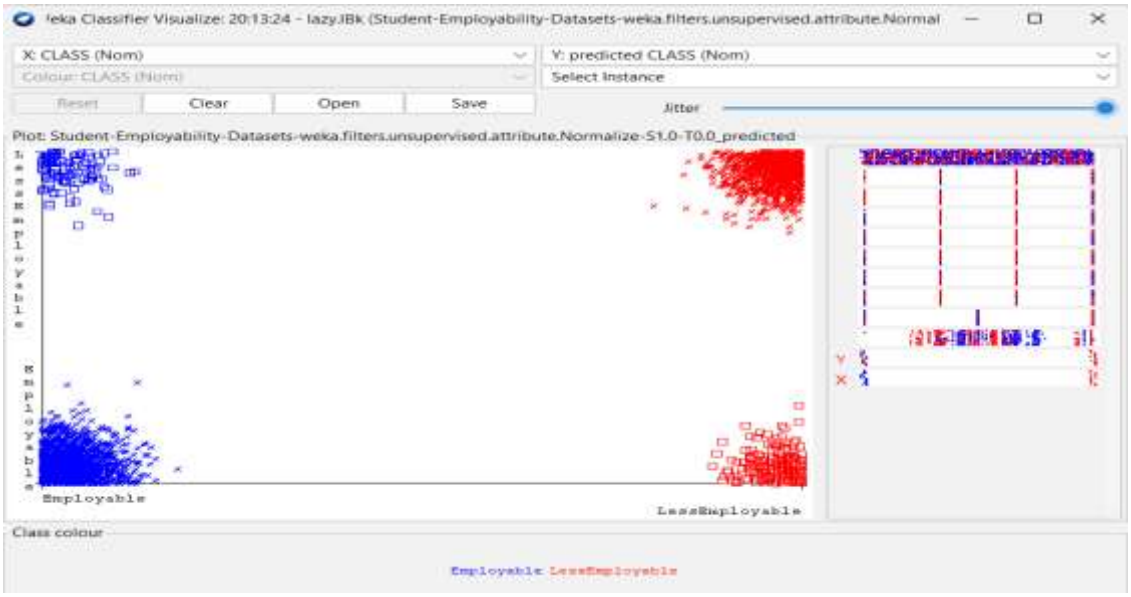


Figure 7 is the WEKA visualization of the IBk classifier with actual class values on the x-axis and predicted class values on the y-axis. The blue points refer to the Employable class, whereas the red points refer to the LessEmployable class. This is shown by the points clustering densely in the correct classes. This is because only a few are scattered outside the dominant clusters, indicating borderline or misclassified cases. In general, the diagram shows that IBk separated the two classes well with some overlapping cases with each other.

Figure 8: Margin Curve Visualization of Logistic Regression

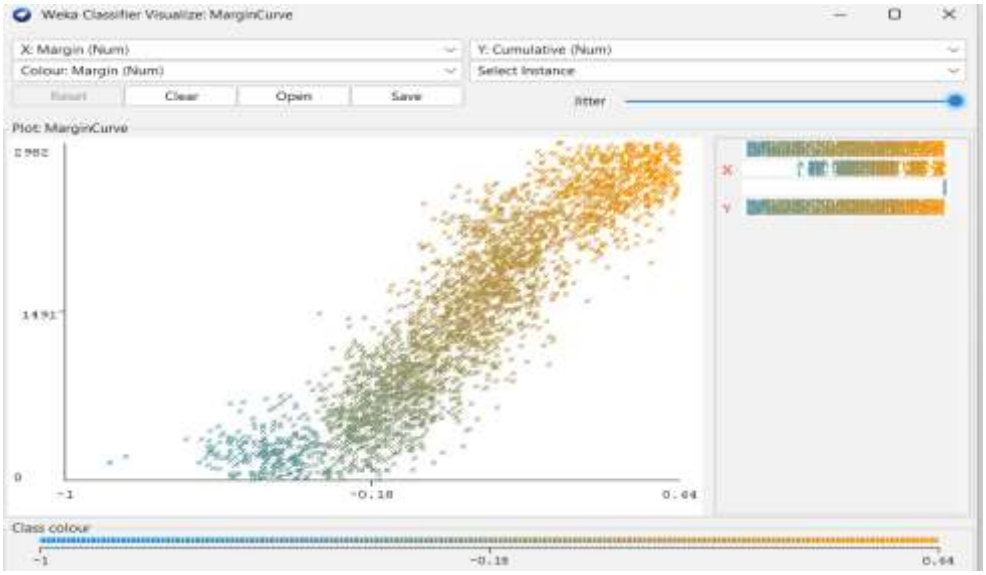


Figure 8 presents a margin curve generated using WEKA when Logistic Regression was selected as the classifier. On the x-axis, the margin value is represented, whereas the cumulative number of instances is denoted on the y-axis. The margin side is orange at the positive end, while at the negative margin, it is blue (Figure 8). Instances with weak or incorrect classification are represented by points on the left, while strong and confident correct classifications are indicated by points on the right. The upward pattern and the location of points mostly in the positive margin region indicate that Logistic Regression classified many instances with accepted confidence levels, although some cases were close to the negative and central margin areas, indicating borderline or misclassified cases.

Figure 9: Naïve Bayes Results

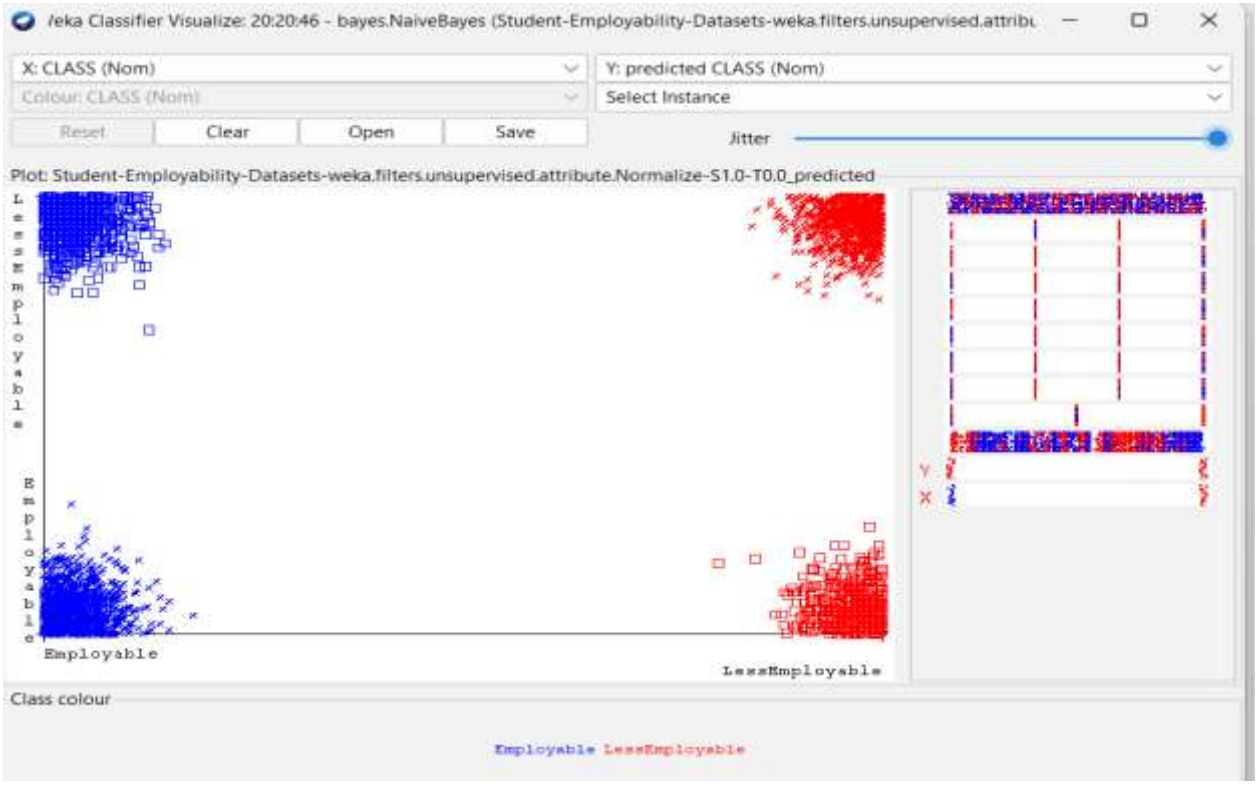


Figure 9 depicts the WEKA classification plot for the Naïve Bayes algorithm. The Naïve Bayes classification plot (Figure 9) in WEKA uses blue points to mark Employable students and red points to mark LessEmployable students. The plot of the WEKA Naïve Bayes classification (Figure 9) shows That Employable students are in blue points and LessEmployable students are in red points, and similar to IBk plot, the pattern ensues in two major clusters, but several points are dispersed more widely around the class boundaries. In other words, the two classes could be distinguished, at least to some extent, in terms of overlapping actual and predicted outcomes. The diagram suggests that the model was good at classifying the obvious cases but had some problems when dealing with problematic cases with varying but closely related values of the interview attributes.

Interpretation of Results in the Context of Objectives

The visual results are analogous to the primary aim of this study, which is to compare the chosen WEKA classifiers for employability prediction. The visual results are in line with the overall central goal of the study which is "A comparison of selected WEKA classifiers for employability prediction". For example, here Naïve Bayes seems to have less of clusters, unlike the IBk and Logistic Regression classifiers, while

the margin curve indicates that negative and positive margins were achieved by many classifications, leaning towards positive margins more (Figure 4: visual statistical results). The figures in this section combined with results in Table 3 provide a further basis of concluding that WEKA classifiers can be used as dependable employability prediction tools, with SMO still preferred as the best performing model in relation to overall accuracy.

9 Key Findings

1. SMO is the Superior Classifier: For this particular dataset of mock interview evaluations, the most accurate predictive capability has been delivered by Support Vector Machines (SMO). Outperforming distance-based (IBk) predictors, statistical (Logistic Regression), and probabilistic (Naïve Bayes) models.
2. Reliability of Distance-Based Learning: Although outperformed by SMO, the IBk classifier demonstrated proven reliability. Compared to SMO with outperformed results, the IBk classifier was proven to be highly reliable when making predictions with nearly 90% accuracy coupled with a low mean absolute error (0.1086) proving that similarly profiled students based on their soft skillsets present highly predictable employability outcomes.
3. Soft Skills as Primary Predictors: The data confirms that technical knowledge is not enough. Two critical variables that discriminate between employable and less employable students are 'Ability to Present Ideas', and 'Self-Confidence' (ibid).

10 Implications

The findings have great practical implications for Higher Education Institutions. By feeding mock interview rubrics into a SMO-based WEKA model, university placement offices can generate automated, highly accurate "employability scores" on forthcoming graduates. This is meant to institutionalize the subjective matter of interview feedback into a measurable indicator. Institutions can identify the "Less Employable" Students much earlier (even months before graduation), this would allow the institutions necessary time to implement targeted interventions, specialized communication training workshops, and confidence boosting seminars.

11 Limitations

Although the results are highly accurate, the study exclusively uses data from mock interviews. Therefore, the perception of an evaluator concerning the employability of a student in a mock interview does not necessarily imply that the student will be hired by a corporation. Excluding academic variables, such as the degree program and technical exam results, guarantees that the model predicts only the presentation bit of the employability factor and not the technical competency that an employer can expect from certain industries.

12 Conclusion

The four research questions in this study were answered by showing that SMO, with 92.54% and outperforming IBk, Logistic Regression, and Naïve Bayes, is the data mining algorithm with the highest accuracy in predicting Filipino students' employability. The key attributes influencing employability in terms of soft-skill framework, which differentiated Employable from LessEmployable students remained consistent throughout the dataset, are namely, Mental Alertness, Communication Skills, Self-Confidence,

and Ability to Present Ideas. On the basis of comparison of WEKA classifiers, SMO performed best in terms of overall predictive performance, but offered less interpretability, IBk was the best and reliable similarity-based model, Logistic Regression provided moderate-to-high degree of accuracy, while offering enhanced interpretability, and the Na¨ıve Bayes served as a baseline model and produced the lowest level of accuracy. On the basis of these results, the study recommends data-driven strategies for universities in the Davao Region include screening for mock interview assessment, employing the use of employability scores for early identification of students at risk of being unemployable, and instituting targeted interventions in communication, confidence-building, mental alertness, and presentation skills based on SMO. Overall, the study confirms ML can be a practical and objective support tool for improving employability preparation for Filipino students.

13 Recommendations

The recommendations below are given based on the conclusions drawn.

1. Integration in Career Centres: Universities in the Davao Region should integrate WEKA-based (SMO) predictive modelling into career centres to be applied to evaluating students after conducting a mock interview.
2. Curriculum Improvement: HEIs should institutionalize dedicated courses on the highest-weighted attributes in this study: Communication Skills, Mental Alertness, and the Ability to put across one's ideas.
3. Future Research: Future studies should try to combine this soft-skills dataset with hard academic data (GPA, technical test scores) and also keep track of the actual employment status of the student's post-graduation so that the mock-interview predictions can be compared to real-life hiring outcomes.

14 References

1. Casuat, C. D., & Festijo, E. D. (2024). Predicting Engineering Students' Employability Using Data Mining Classification Techniques. 2024 IEEE 3rd Conference on Information.
2. Moumen, A., El Bakkouri, I., Kadimi, H., Zahi, A., Sardi, I., Tebaa, M., Bousserrhine, Z. & Baraka, H. (2022). Machine Learning for Students Employability Prediction. Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning (BML), 274-278.
3. Olipas, C. N. P. (2026). Predictive modeling and explainability of student employability in the Philippines using Random Forest and Shapley Additive Explanations. Interdisciplinary Journal of Information, Knowledge, and Management, 21, Article 2. https://doi.org/10.28945/5690
4. Almalawi, A., Soh, B., Li, A., & Samra, H. (2024). Predictive models for educational purposes: A systematic review. Big Data and Cognitive Computing, 8(12), 187. https://doi.org/10.3390/bdcc8120187
5. Saidani, M., et al. (2024). Prediction of Student Employability through Internship based on Big Data Analysis. Journal of Electrical Systems, 20(3s), 2749-2761.
6. Moutni, A. (2023). Students' employability dataset – Philippines Data set. Kaggle. https://www.kaggle.com/datasets/anashamoutni/students-employability-dataset