Philippine Journal of Interdisciplinary Research (PJIR)
Volume I, Issue 1 (January 2025)

Predicting Employability of CCA ICSLIS Students Using Ensemble Methods

Harwin C. Mendoza[1]

[1]Institute of Computing Studies and Library Information Science, City College of Angeles, Angeles City, Pampanga

Corresponding email: harwinmendoza@cca.edu.ph
Received: 28 Mar 2023; Accepted 11 May 2023; Available online: 08 October 2024

Abstract. Education helps graduates find jobs because it is one of the most important ways to improve their skills. The degree of difficulty an individual faces in obtaining a degree may appear to be a clear path to success, and the by-product, a diploma, as the 'end-all, be-all,' but popular opinion plainly reveals that it is only a head start in the employment life. In this study, the accuracy of various machine learning methods was compared to develop an ensemble prediction model with the ability to predict the employability of ICSLIS graduates by making substantial use of data mining techniques. This was determined based on evaluation metrics and an analysis of feature importance based on responses from ICSLIS graduates, indicating that employability was strongly correlated with technical skills, having a portfolio connected to their degree or major, and technical certificates. As a result, the information acquired can be used in the formulation of a wide variety of policies, programs, and strategies intended to increase the job opportunities available to students.

Keywords: *IT employability, Binary Classification, Student Competency, Ensemble Methods*

INTRODUCTION

Information technology (IT) is constantly evolving, and employers seek highly skilled professionals who can meet the demands of the industry. However, there is growing concern that many IT graduates lack the necessary skills and experience to be successful in the workforce. Therefore, there is a need to develop more accurate and reliable predictors of IT students' employability. This literature review aims to identify and analyze existing studies that have examined the factors that contribute to employability among IT students and the methods used to predict their employability. (Patel A. et al., 2020).

Several studies have identified factors that contribute to employability among IT students. These factors can be broadly classified into three categories: technical, experience, and soft skills. Technical skills refer to students' knowledge of programming languages, databases, software development, and other technical areas. Experience includes internships, co-op programs, and other work-related experiences that provide students with hands-on experiences in the field. Soft skills refer to a student's ability to communicate effectively, work

in teams, and solve problems (Thakar P. et al., 2017).

Several studies have found that technical skills are important for employability in the information technology (IT) industry. For example, Wang et al. (2016) found that programming skills are the most important factor for employability among IT students. Similarly, Alshammari and Al-Qirim (2015) found that technical skills are key predictors of employability in the IT industry.

Experience has also been identified as a significant factor in IT students' employability. For example, Khazaei et al. (2016) found that work experience was positively associated with employability among IT students. Additionally, Sardar and Al-Saif (2017) found that internships and other work-related experiences are positively associated with employability among IT graduates.

Soft skills are also been shown to be important for employability in the IT industry. For instance, Alghamdi et al. (2018) found that communication skills and teamwork are significant predictors of employability among IT students. Similarly, Chua et al. (2018) found that problem-solving skills were positively associated with employability in the IT industry.

Various methods have been used to predict the employability of IT students, including statistical models, machine-learning algorithms, and ensemble methods. Mokhtari et al. (2020) used a random forest algorithm to predict IT students' employability. The study found that the algorithm had a high accuracy rate in predicting IT students' employability. Another study by Liu et al. (2017) used a support vector machine (SVM) to predict the employability of IT graduates. The study found that SVM had a higher accuracy rate than logistic regression in predicting the employability of IT graduates (Alghamlas M. and Alabduljabbar R., 2019).

Ensemble methods, which combine multiple models to make a more accurate prediction, have also been used to predict IT students' employability. For example, Sabri et al. (2018) used an ensemble method to predict IT graduates' employability. The study found that the ensemble method had a higher accuracy rate than individual models in predicting IT graduates (Songpan W., 2020).

Objectives of the Study

The primary aim of this research was to develop a machine-learning-based forecasting system for predicting the employability of students from the ICSLIS of the City College of Angeles. The intention was to provide actionable insights and guidance to the ICSLIS department to effectively support undergraduates in securing employment opportunities. Additionally, the study sought to achieve the following objectives:

1. To develop a system that can provide a binary classification about the student's employability.
2. To test the system's accuracy.
3. To determine which among the three machine learning techniques used had the highest accuracy.

Scope of the Study

In this binary classification problem, the task is to predict whether a student is "employable" or if he/she "needs intervention" based on a set of features. The features include the year of graduation, study program, thesis grade, organization affiliation, Latin honors, awards received, and involvement in extracurricular activities. The dataset consists of 196 training samples. To evaluate the performance of the model, the dataset was divided into a 70% training set and a 30% test set, allowing for training and testing the model on separate data subsets. The goal is to develop a machine-learning model that can accurately classify students into appropriate categories of employability based on the features provided. The software utilized for this project included Python, NumPy, Pandas, and Scikit-learn. Python serves as the programming language for development, whereas NumPy provides powerful numerical computing capabilities. Pandas facilitates efficient data manipulation and analysis, and scikit-learn offers a comprehensive set of machine learning algorithms and tools for model training and evaluation.

Delimitation

This study is subject to certain limitations, owing to time constraints and data availability. The time constraint restricts the extent to which data can be collected, processed, and analyzed, potentially affecting the depth and comprehensiveness of the findings. Additionally, the absence of a school system limits the available data, potentially limiting the scope and richness of the information that can be obtained. These limitations may affect the generalizability and completeness of the employability prediction model because it is based on a subset of data and may not capture all relevant factors. It is important to acknowledge these limitations when interpreting and applying this study’s findings.

METHOD

The agile methodology was adopted as the framework for developing a binary classification model to predict the IT employability of CCA ICSLIS students. Iterative planning was conducted to break down the development process into manageable iterations, focusing on specific subsets of features and data requirements for each iteration.

The researcher collected relevant data from various sources, such as student records and academic databases, and preprocessed the data by handling missing values and standardizing formats. Feature engineering was employed to extract meaningful features from the collected data, considering factors such as academic performance, skills, internships, and extracurricular activities that influence IT employability. For model training and evaluation, the researcher selected an appropriate machine-learning algorithm for binary classification and split the dataset into training and testing sets. The model was trained using the training data and evaluated using evaluation metrics such as accuracy, precision, recall, and F1 score. Through an iterative development approach, the researcher incorporated feedback from stakeholders to refine the model, adjust features, fine-tune hyperparameters, and explore different algorithms to improve its accuracy and performance. Once the final

model was developed, the researcher deployed it and validated its predictions using unseen data from the CCA ICSLIS students. Continuous monitoring and evaluation of the model's performance are conducted with ongoing collaboration and feedback from stakeholders to ensure its effectiveness and relevance.

Throughout the research process, the researcher embraced the principles of agile, fostering communication, collaboration, and adaptability. This approach enables the researcher to develop a robust binary classification model for predicting IT employability and addressing the specific needs and challenges of CCA ICSLIS students.

RESULTS

In this section, the researcher explores a dataset containing information about students gathered from the Institute of Computing Studies and Library Information Science. Following the completion of all the preprocessing steps, one hot encoding technique was performed wherein the categorical data were then transformed into numeric data, which maintains the sequence of the data during the conversion process.

The data were divided into two categories: training and testing. The dataset includes information such as the student's year of graduation, program of study, capstone grade, organizational involvement, competition involvement, Latin Honor, and awards received. The researchers used a voting algorithm consisting of logistic regression, random forest, and K-nearest Neighbor models. A voting algorithm is a highly efficient ensemble-learning technique that integrates the predictions of different models to obtain a final result. By leveraging the strengths of each individual model, this study aims to build a robust and accurate predictor of whether a student will be employed after graduation. Through the researcher's exploration and analysis of the data, the trends and insights that helped the researcher build an effective model were uncovered. Tables and charts below explain the essential details of the dataset.

Name	Year	program_of_study	employability	Capstone	Organization I	Competiti	with Latin	Awards
1	2020	1	Employed	2	1	1	1	1
2	2020	1	Unemployed	2.5	2	1	1	1
3	2020	1	Employed	1	1	2	1	1
4	2020	1	Employed	2	1	1	1	1
5	2020	1	Employed	2.25	1	1	1	1
6	2020	1	Employed	2.5	1	1	1	1
7	2020	1	Employed	3	1	1	1	1
8	2020	1	Employed	1.25	1	1	1	1
9	2020	1	Employed	2	2	1	1	1
10	2020	1	Employed	1.25	1	1	1	1
11	2020	1	Unemployed	2.5	2	1	1	1

Figure 1. *Portion of the Dataset*

Figure 1 shows a portion of the processed and cleaned datasets that can be used for further analysis by the researcher. The table also shows the features remaining after removing unnecessary features from the uncleaned dataset.

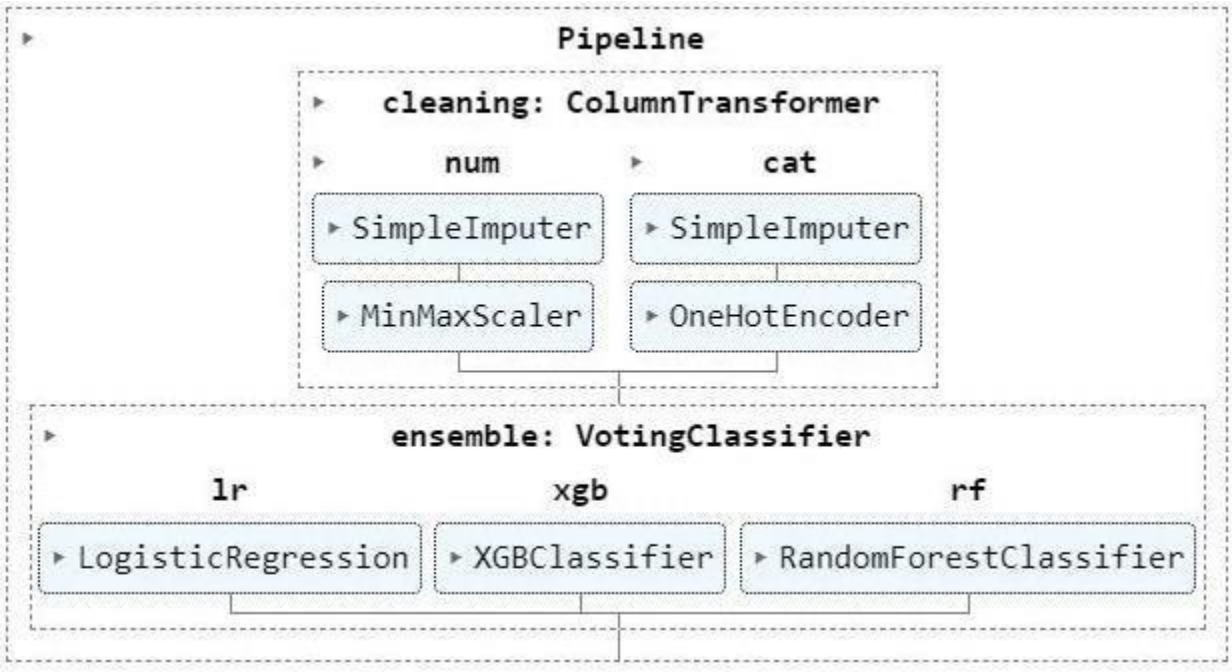

Figure 2. *Pipeline Used in the Model*

In the modeling part of the analysis, several machine-learning algorithms were utilized to build a robust predictive model. Specifically, Random Forest, K-Nearest Neighbors, and Logistic Regression algorithms were used to train and test the model and combine them using a voting algorithm. The Voting algorithm helped the researcher aggregate the results of the individual models and generate a final prediction that is more accurate and stable than any individual model. This is because the Voting algorithm considers the strengths and weaknesses of each individual model and leverages their respective predictive abilities to generate a more robust and accurate result. Overall, using a voting algorithm is a powerful way to ensure that the model is accurate and reliable, and leverages the best of each individual model to generate a final prediction.

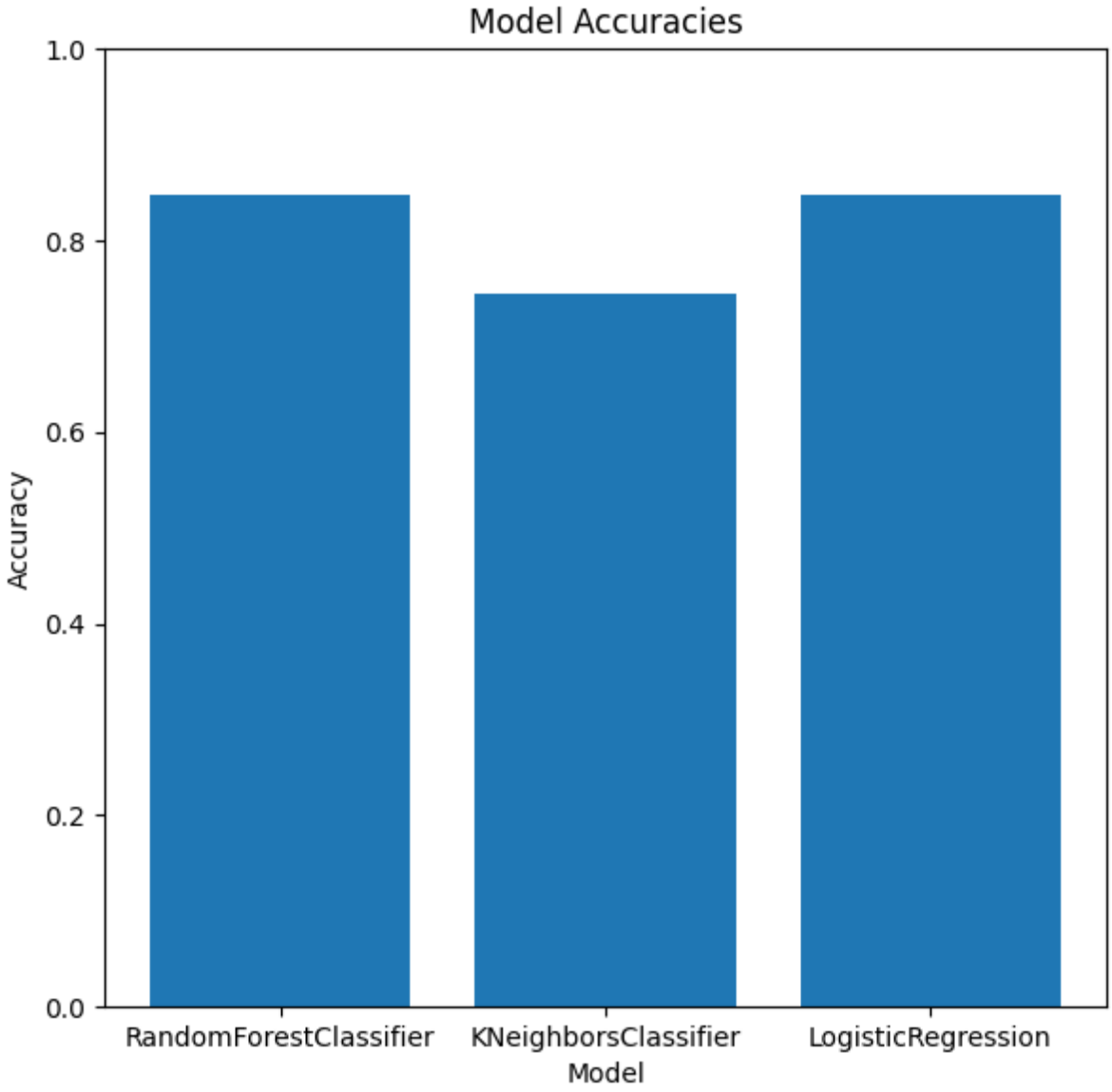

Figure 3. *Model Accuracies*

Model: RandomForestClassifier

	precision	recall	f1-score	support
Employed	0.86	0.96	0.91	46
Unemployed	0.75	0.46	0.57	13
accuracy			0.85	59
macro avg	0.81	0.71	0.74	59
weighted avg	0.84	0.85	0.83	59

Figure 4. *Random Forest Classifier Accuracy*

Model: KNeighborsClassifier

	precision	recall	f1-score	support
Employed	0.77	0.96	0.85	46
Unemployed	0.00	0.00	0.00	13
accuracy			0.75	59
macro avg	0.39	0.48	0.43	59
weighted avg	0.60	0.75	0.67	59

Figure 5. *K-Nearest Neighbors Classifier Accuracy*

Model: LogisticRegression

	precision	recall	f1-score	support
Employed	0.84	1.00	0.91	46
Unemployed	1.00	0.31	0.47	13
accuracy			0.85	59
macro avg	0.92	0.65	0.69	59
weighted avg	0.87	0.85	0.81	59

Figure 6. *Logistic Regression Accuracy*

Based on the performance metrics of the models, two out of three models have high precision scores, with Random Forest having 81% and Logistic Regression having 92%, respectively. However, when it comes to recall and F1 score, Random Forest outperformed the other two models with a recall of 71% and an f1-score of 74%. The K-nearest Neighbor had the lowest scores across all metrics.

DISCUSSION

The employability of individuals is intricately linked to the educational offerings of universities, as they significantly influence the career progression and academic development of their graduates. It is incumbent on universities to equip their students with the necessary skills and knowledge to thrive in their chosen professions, thereby addressing the evolving demands of industries and remaining competitive.

The quality of education a candidate has received and their level of preparedness for the demands of a job are paramount in determining their suitability for employment. Enhancing one's employability prospects is best achieved by acquiring robust education and establishing a solid foundation. Ultimately, a bachelor's degree underscores the potential employers for which an individual has received comprehensive education. Additionally, it attests to a person's possession of fundamental skills, such as effective communication, comprehension,

and a foundational skill set. These proficiencies are integral components of nearly all bachelor's degree programs, and showcasing competence in these areas marks a significant step towards bolstering overall employability.

This investigation effectively identified the elements contributing to a precise forecast of graduates' employability within the realm of the ICSLIS. The researcher successfully devised a voting ensemble model that yielded employment predictions based on these variables. By amalgamating diverse models, ensemble learning enhances the efficacy of machine-learning outcomes. Compared with relying on a solitary model, this approach facilitates the generation of a markedly superior predictive performance. The fundamental principle involves identifying a cohort of classifiers that then collaboratively cast their votes. A Voting Classifier, a variant of the machine learning algorithm, undergoes training on an assemblage of multiple models and subsequently makes predictions by considering the highest probability choice among the models for each category as the outcome. It amalgamates the results of the individual classifiers fed into it, and focuses its prediction of the output category on the classification that garners the most substantial majority of votes.

In lieu of constructing discrete standalone models and assessing their individual accuracies, the underlying concept is to formulate a single model that learns from these constituent models and predicts outcomes based on the cumulative majority consensus for each output category. This circumvents the need to create entirely separate models and the subsequent assessment of their accuracy.

Conclusion

The result of the ensemble is often determined by applying various voting algorithms to the data. There are a number of voting techniques, and the outcomes they produce may vary owing to a number of factors, such as the algorithm types that are used. In many instances, the performance of the ensemble-based classifier was superior to that of individual classifiers. The primary objective of this research is to demonstrate that a voting classifier can recognize student employability. The Voting ensemble classification used was a combination of the Random Forest Classifier, K-Nearest Neighbor and Logistic Regression Classifier. It was used to address the prediction model discussed in this study.

As mentioned, one of the characteristics that strongly influenced the prediction of employability was Technical Skills. This is because every work has a particular set of skill requirements; thus, technical skills are usually among the most sought-after and essential for any profession for many different reasons. Technical Skills provide students with underlying knowledge and expertise, as well as the potential to help them work more efficiently, enhance their confidence, and make them more valuable candidates for prospective employers. This is because employers anticipate working with qualified teams, in which they may have confidence in achieving the results they require. In this case, employers continually look for candidates with technical skills to fill open positions in their team. Similarly, having a portfolio related to the graduate’s course/major allows the graduate to demonstrate their abilities, establish their reputation, and grow in self-branding. Throughout the process of job

interviews, graduates may market themselves and demonstrate all the work they have done by referring to their portfolios as a tool.

Despite the significant advances made as a result of this research, there are a few limitations that must be considered. Initially, there was a shortage of data that could be used to accurately describe the employability of ICSLIS graduates; therefore, the research relied on data collected from the Faculty in Charge. In addition, the researcher reached other ICSLIS graduates via existing Facebook groups, but there was a high non-response rate among graduates in the ICSLIS. It was difficult to obtain a large number of individuals to complete the survey and obtain feedback. Despite this, the study survey produced complete data collection that contained 195 records. Consequently, the findings of this study need to be interpreted with regard to its limitations. Consequently, the researcher recommends expanding the dataset size for any future research that might be conducted to ensure that the model will have a higher level of accuracy. When predicting the employability of graduates, further research should focus on other fields of study. Future research may integrate a decision support system as a prescriptive solution with the aim of lowering the unemployment rate in ICSLIS.

REFERENCES

Alghamlas, M., & Alabduljabbar, R. (2019). *Predicting the suitability of IT students' skills for the recruitment in Saudi labor market*. https://ur.booksc.me/book/76520761/af15ca

Bharambe, Y., Mored, N., Mulchandani, M., Shankarmani, R., & Shinde, S. G. (2017, September). Assessing employability of students using data mining techniques. In *2017 international conference on advances in computing, communications and informatics (icacci)* (pp. 2110-2114). IEEE.

Casuat, C. D., Festijo, E. D., & Alon, A. S. (2020). Predicting students' employability using support vector machine: a smote-optimized machine learning system. *International Journal*, *8*(5), 2101-2106.

García-Peñalvo, F., Cruz-Benito, J., Martín-González, M., Vázquez-Ingelmo, A., Sánchez-Prieto, J. C., & Therón, R. (2018). Proposing a machine learning approach to analyze and predict employment and its factors.

Mezhoudi, N., Alghamdi, R., Aljunaid, R., Krichna, G., & Düştegör, D. (2023). Employability prediction: a survey of current approaches, research challenges and applications. *Journal of Ambient Intelligence and Humanized Computing*, *14*(3), 1489-1505.

Patel, A., Mascarenhas, S., Thomas, A., & Varghese, D. (2020, June). Student performance analysis and prediction of employable domains using machine learning. In *Proceedings of the International Conference on Recent Advances in Computational Techniques (IC-RACT)*.

Rojarath, A., & Songpan, W. (2020). Probability-weighted voting ensemble learning for classification model. *Journal of Advances in Information Technology Vol*, *11*(4).

Römgens, I., Scoupe, R., & Beausaert, S. (2020). Unraveling the concept of employability, bringing together research on employability in higher education and the workplace. *Studies in Higher Education*, *45*(12), 2588-2603.

Thakar, P., & Mehta, A. (2017). A unified model of clustering and classification to improve students' employability prediction. *International Journal of Intelligent Systems and Applications*, *9*(9), 10.

Verecio, R. L. (2018). Predicting employability skills among information technology graduates of philippine state university in their on-the-job training using J48 algorithm. *Indian Journal of Science and Technology*, *11*(37), 130842.

Vinutha, K., & Yogisha, H. K. (2020). Employability Prediction of Engineering Graduates using Machine Learning Algorithms. *International Journal of Recent Technology and Engineering (IJRTE)*, *8*(5), 2277-3878.