UMYU Scientifica, Vol. 2 NO. 1, March, 2023, Pp 001 – 009
https://doi.org/10.56919/usci.2123_001
ISSN: 2955 – 1145 (print); 2955 – 1153 (online)
A periodical of the Faculty of Natural and Applied Sciences, UMYU, Katsina

ORIGINAL RESEARCH ARTICLE

Machine Learning for Predicting Students' Employability

Muhammad Hadiza Baffa[1*], Muhammad Abubakar Miyim[2], Abdullahi Sani Dauda[3]

[1]Department of Computer Science, Federal University Dutse, Nigeria.
[2]Department of Information Technology, Federal University Dutse, Nigeria.
[3]Department of Computer Science, Federal University of Kashere, Gombe, Nigeria.

ABSTRACT

Graduates' employability becomes one of the performance indicators for higher educational institutions (HEIs) because the number of graduates produced every year from higher educational institutions continues to grow and as competition to secure good jobs increases, it is significant for HEIs to understand the employability of graduates upon graduation and highlight the reasons. To predict students' employability before graduation, machine learning models were employed. These include logistic regression; decision tree, random forest, and an unsupervised clustering (K-Means) algorithm. This research, therefore, aims to predict the full-time employability of undergraduate students based on academic and experience employability attributes – including cumulative grade point average (CGPA), student industrial work experience scheme (SIWES), co-curricular activities, gender, and union groupings before graduation. Primary datasets of 218 graduate students in the last four academic calendar years (2016 – 2021) from the Computer Science Department of Federal University Dutse were rated. The results demonstrate that Random Forest Classifier predict students employability the best with an accuracy of 98% and f1-score of 0.99 compare to logistic regression and decision tree. Furthermore, using more students' data with more attributes including academics and extracurricular activities can improve the models performance and predict students' employability.

ARTICLE HISTORY

Received December 20, 2022
Accepted January 23, 2023
Published February 13, 2023

KEYWORDS

Data mining, Machine learning, Employability, Prediction

INTRODUCTION

Higher Educational Institutions (HEIs) play vital role in the economic development of any country in which Nigeria is not isolation. It is an industry that helps in supporting the rest of the industries by giving them trained workforce to boost their productivity that eventually improves the GDP of the country. Previously, HEIs did not consider employability as an important concern and lacking in advising students in selecting courses that are more beneficial to their development. Nowadays, employability of students has become a great concern for HEIs as most of them have been considering early prediction of students' employability as a desired action that is continuously needed for decision taking. As competition for jobs increases, students must identify academically and experiential opportunities to differentiate themselves in the entry-level labour market (Hugo, 2018). Moreover, as universities are increasingly held accountable for student outcomes and continuously amass student data, the data can serve to inform which students are more likely to be employed upon graduation and the reason. Though, educational institutions becoming more employment-oriented, employment of graduate's students from many institutions become a major issue in constructing the standing of the institution and therefore a major concern. As higher educational institutions gradually take part to attract and keep students in their institutions, these may take the benefit of data mining, mostly in predicting enrollment (Mishra *et al.*, 2017; Wanjau & Muketha, 2018).

It is evident that every year, millions of secondary school students sit for Unified Tertiary Matriculation Examination (UTME) conducted by Joint Admission and Matriculation Board (JAMB) and millions of them end up not gaining admission into universities and other tertiary institutions. Furthermore, our universities, on the other

Correspondence: Muhammad H. B., Department of Computer Science, Federal University Dutse, Nigeria. ✉ svelte4real@gmail.com , Phone +234 813 900 2711
How to cite: Muhammad H. B., Muhammad A. M., Abdullahi S. D. (2023). Machine Learning for Predicting Students' Employability. UMYU Scientifica, 2(1), 241 – 253. https://doi.org/10.56919/usci.2123_001

hand, keep producing thousands of graduates in various disciplines into the labor market. Such trend is becoming a serious concern as every student struggles for a white-collar job after graduation which the labor market cannot accommodate all.

Higher Educational Institutions do give great concern to employability of their Students and is considered vital for the institutions as it is often used as a metric for their achievement (Mezhoudi *et al.*, 2021). The disappointment of landing a job for university students might cause severe social penalties such as intoxication and self-destruction. In addition to academic performance, lifeless biases can develop one key problem in pursuing the jobs for student graduates. Therefore, it is essential to apprehend these biases so that it can help the students at the initial step with more personalized interference (Guo *et al.*, 2020).

Machine learning techniques are typically categorized into supervised and unsupervised techniques with the supervised machine learning starting from previous knowledge of the desired outcome in the form of labeled data sets that monitor the process of training, while unsupervised machine learning works directly on unlabeled data. In the case of lack of labels to orientate the learning process, these labels need to be "discovered" by the learning algorithm (Palacio-Niño & Berzal, 2019).

Much research has been done about Graduates' employability using various machine learning algorithms to show how the performances are suitable for HEIs. Human Resource units are required to sort or group the applicants' profiles manually before can select the appropriate candidate for the position they are recruiting. This, however, required a considerable amount of time, this is even dependent on the number of applications (Oladokun *et al.*, 2008).

Authors in Hugo, (2018) proposed a study for determining the extent undergraduate student academic performance (CGPA) and student work experience (SIWES), as well as co-curricular activities, helps predict student chances in securing employment. Their research methodology adopted known and innovative machine learning models (Logistic Regression, Decision Tree, Deterministic Analysis and Neural Network) to predict employment before graduation. The results revealed that employment before graduation can be forecasted using neural network as the most exact predictive model with 73% accuracy.

In the work of Kumar & Babu, (2019), a sample data of about 500 students were collected in various engineering colleges within Hyderabad. The researchers used supervised machine learning techniques (Support Vector Machines, Naïve Bayes, Decision Tree and K-Nearest Neighbor) in predicting students' employability and to regulate the factor impacting the employability of students. The outcome of their work indicated Decision Tree and Support Vector Machines were better than Naïve Bayes and K-Nearest Neighbor in predicting the employability of students with 98% accuracy.

It is observed that factors, like communication, aptitude and reasoning skills, mentorship, family status, quality of teaching in university, etc., have significant impact on student academic performance. In Philippines, authors in Esquivel & Esquivel, (2020) proposed a study on several features of freshman aspirants in upsetting admission status in universities. The predictive model used was logistic regression (LR) established to assess the possibility of candidates gaining admission into a university or not. The results showed that, given inadequate information about potential students HEI can implement a technique to complement management decisions and offer an estimate of class size, allowing the institution to enhance the allocation of assets and have control better over tuition. In another scenario, Authors proposed the work employability of graduate students based on their performance in academics and skills and applied different machine learning models (Vinutha & Yogisha, 2021). The ANN algorithm found to have satisfactorily performed in terms of accuracy of 87.42%, than LR and Naïve Bayes with 85.2%, and 84.21% respectively.

Supervised and unsupervised machine learning classifiers were both considered in the work of Oladokun *et al.*, (2008) where Naïve Bayes, LR, SVM, RF and DT were proposed as the machine learning algorithms for classification and predicting aspirants' attributes for a job to meet the choice norms for an industry, while considering their academic performance. Though all fared well in the predicting the accuracy, however, the result appear to favour LR as the best performing algorithm among them. The datasets considered were not enough to test other features and therefore suggested that future work should include NN to be used for the prediction.

Furthermore, in Mezhoudi *et al.*, (2021), a proposed Employability prediction was focused on the review of methodologies by recognizing the important factors touching employability that can extremely assist all participants. Knowing their flaws and assets, students might have a plan for their occupation. As the number of graduates produced every year from higher educational institutions continues to grow, hence, it is significant to highlight the employability of graduates that delivers a broad roadmap, permitting the application of data mining for employability.

The objective of this study is to design student employability predictive models using academic and extracurricular activities attributes like CGPA, SIWES

result, and place of student industrial work experience scheme (SIWES), gender, union group, and year of graduation as the determining factors in the prediction using machine learning classifiers namely; Decision Tree, Logistic Regression, and Random Forest.

The purpose of this study is to determine how undergraduate students' academic and experience employability attributes including CGPA, SIWES, Union Group and Gender can predict chance of getting employment before graduation. The following questions guide this research:

1. To what extent can student academic and experience employability signals predict student employability before graduation?
2. Which machine learning method is most accurate in predicting student employability before graduation?
3. What employability signals are most significant in predicting student employability before graduation?

MATERIALS AND METHODS

The proposed methodology is depicted in figure 1. It describes how data was collected for this study, preprocessing steps to clean the dataset, including handling missing values, eliminating repetitive or extraneous variables, and encoding and simplifying data. It also defines the development of unsupervised clustering machine learning models to predict employment after graduation as well as describes the evaluation metrics used.

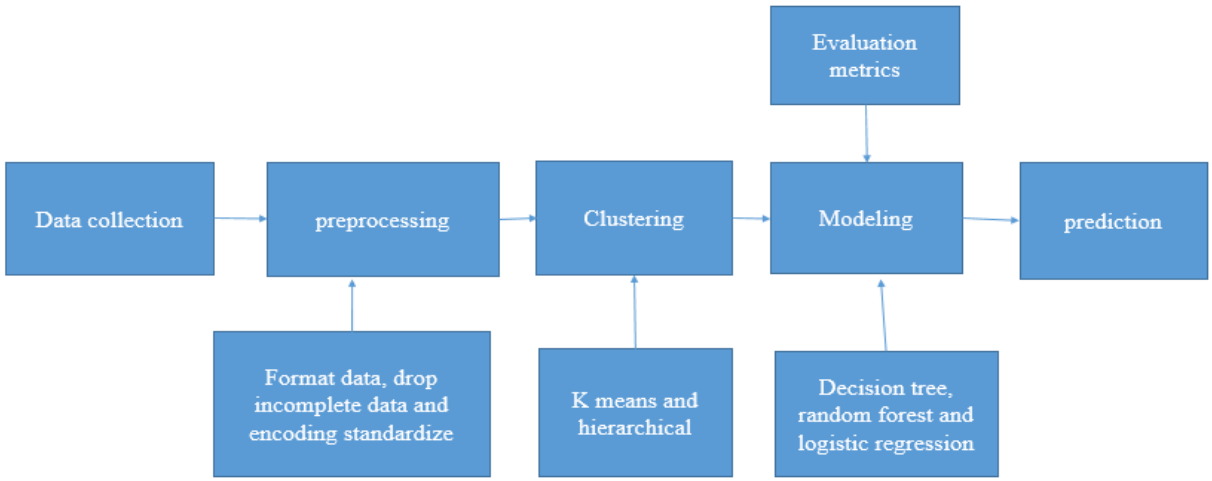

Figure 1: General Methodology of the Proposed System

The software includes python programming language[1] through Pandas library[2] for importing the data in spreadsheet form NumPy library[3] for conducting numerical analysis on the data matplotlib[4], and seaborn[5] for rendering visualizations, then the two algorithms were implemented by importing their various libraries in jupyter notebook[6].

a. Data Collection

The dataset used was collected from Computer Science Department in the Faculty of Computing, Federal University Dutse. The dataset consists of six variables from 218 students' academic performance records that included CGPA, SIWES results, Union Group, gender, matric number, graduation year from four academic years, 2016 to 2021. After collecting the data, all personally identifiable information was removed before converting the non-numeric data into numeric: Gender (male represented as 1, female represented as 2), Union Group (official represented as 1, member represented as 2) respectively.

b. Data Pre-Processing

Pre-processing steps was introduced to the original dataset that made it suitable for analysis, while missing values and extra variables were deleted from the dataset.

[1] https://www.python.org/
[2] https://pandas.pydata.org/
[3] https://numpy.org/
[4] https://matplotlib.org/
[5] https://seaborn.pydata.org/
[6] https://jupyter.org/

The description of the processed datasets is presented in table 1.

Table 1: Description of the Processed Datasets

S/N	Variable Name	Data Type & Value	Variable Description	Descriptive Statistics
1	CGPA	Float	The cumulative GPA at graduation	Counts: 218 min: 1.80000. Max: 4.960000. mean:3.499128, std: 0.702852
2	SIWES results	Integer	Results of SIWES program in which students participate in the university	Counts: 218, min: 55,000000, max: 92,000000, mean: 76.174312,, std: 8.105120
3	Gender	Objects [1, 2]	The sex of a student i.e., Male, Female	Counts: 218 Male(1) 184 , female(2) 34,
4	UNION group	Objects [1, 2]	The number of student organizations in which a student participated i.e., Official or Member	Counts: 218 official (1) 50, member (2) 168,
5	Place of SIWES	Objects [1, 2]	Place of SIWES program in which students participate in the university i.e., ICT base or Non ICT base	Counts: 218 ICT (1) 128, Non ICT (2) 90,
6	Year of Graduation	Integer	The date a student Graduated	Counts: 218, Year 2016, 2017, 2018, 2019, to 2021

c. Unsupervised Clustering Algorithms

Unsupervised clustering algorithms are powerful techniques in predictive modeling. They are used to classify and group similar data points in a higher dataset without worrying about the specific outcome. The cluster analysis is then used to categorize data into structures that are more easily understood and manipulated. In this study, we explored two different clustering algorithms for predicting student employability; namely, K-means and Hierarchical clustering. For the formation of clusters, Elbow method was applied to find the optimal number of clusters in K-means while considering only two variables (CGPA and SIWES) that produce data points into 2, 3, and 4 respectively as shown in the figure 2.

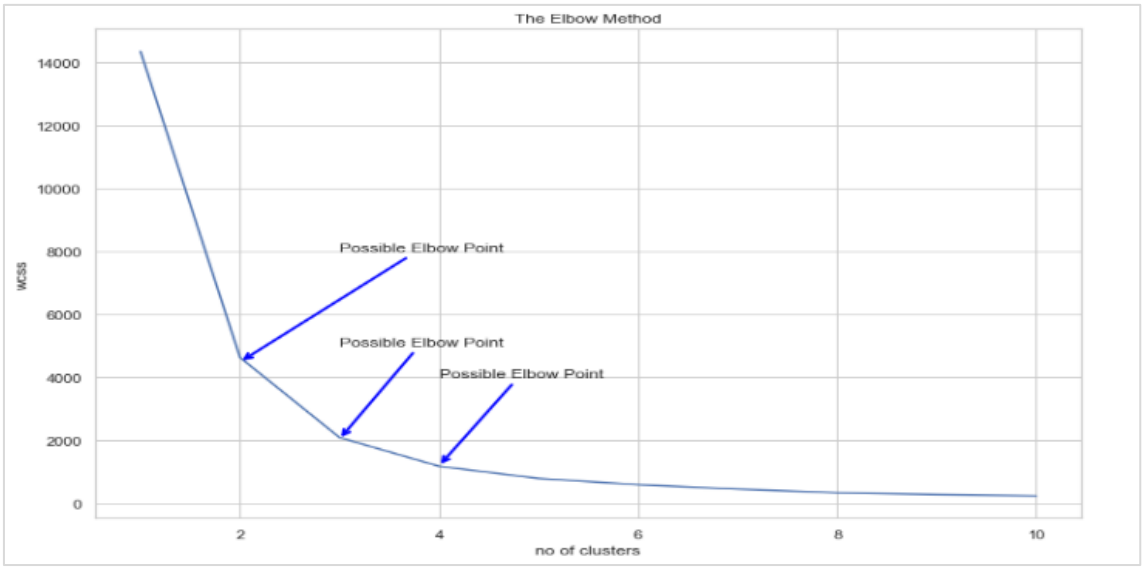

Figure 2: K-means Clustering using Elbow Method

d. Machine Learning Classifiers

After clustering using k means for labeling, the data is then split into 80% meant for training and 20% for validation and setting 3 algorithms (DT, LR and RF) for prediction, their configuration is given in table 2.

Table 2: machine learning classifiers settings

MODEL	PARAMETER SETTINGS
LR	C: 1.0, solver: 'lbfgs', penalty: l2
DT	max_depth: 10, min_samples_leaf: 20, criterion: 'entropy'
RF	n_estimators: 35, max_depth: 5, max_features: 'auto', min_samples_split: 2, min_samples_leaf: 1

e. Evaluation Metrics

In order to understand the performance of the proposed algorithm, some metrics were evaluated and compared with similar works. These metrics used in this study include the following:

i. ACCURACY: most commonly used metrics; it tells how often the classifier is correct in making the prediction and mathematically defined as in Eq. 1.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \ldots\ldots \quad (1)$$

ii. PRECISION: explains how precise our classifier is, it is useful where False Positive (Type I Error) is a higher concern and is expressed mathematically in Eq. 2.

$$Precision = \frac{TP}{TP+FP} \ldots\ldots \quad (2)$$

iii. RECALL: clarifies how many actual positive suitcases were able to forecast properly with our model. It is beneficial where False Negative (Type II Error) is a greater concern. The Mathematical formulation is given in Eq. 3.

$$Recall = \frac{TP}{TP+FN} \ldots\ldots \quad (3)$$

iv. F1 SCORE: this is the vocal mean of the recall and precision and is expressed Mathematically in Eq. 4

$$f1\_Score = 2\frac{precision*Recall}{precision+Recall} \ldots\ldots \quad (4)$$

RESULTS AND DISCUSSION

The student's datasets comprise of four (4) numeric variables and three (3) categorical variables. Only the numerical variables were computed without consideration to the graduation year as its importance is insignificant. In Figure 3 below, based on 2 clusters using CGPA and SIWES result, it indicates that cluster 2 (blue) have a high possibility of students getting employed after graduation due to high marks (75% and above) recorded in SIWES while maintaining their respective CGPA and cluster 1(red) represents students with less possibility of getting employed upon graduation as less than 75% was earned in SIWES. These can be interpreted as having a SIWES score of 75% and above will increase the chance of getting employment irrespective of the students CGPA as the algorithm consider SIWES score more important than the CGPA, while having a good CGPA without performing good in SIWES won't increase the chance of getting employment.

i. Clustering using Kmeans and Hierarchical methods.

In considering 3 and 4 clusters, the results are depicted in figure 4 and 5 respectively. While figure 4 indicates that students with SIWES scores above 80% stand the chance of getting employed before graduating and that of figure 5 shows the likelihood that students scoring average marks of 68% - 75% may be employed after graduation. These can be interpreted as the student possesses skills in the program of their study and has high potential of employability.

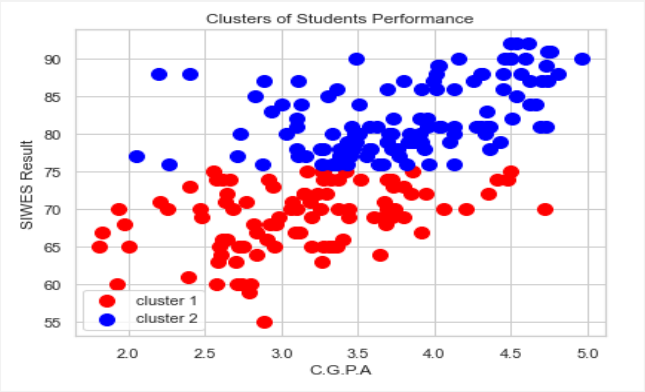

Figure 3: Two Clusters using Two Variables (SIWES and CGPA)

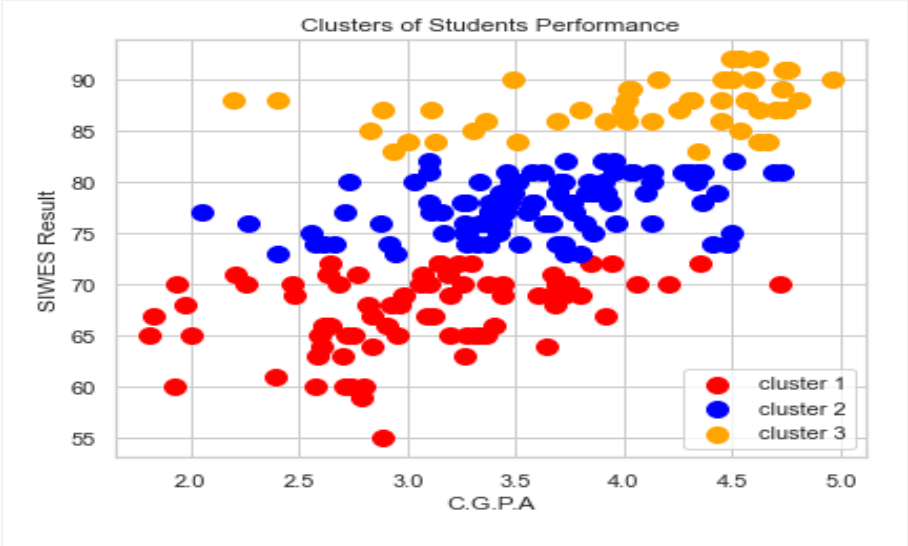

Figure 4 Three clusters using 2 variables (SIWES and CGPA)

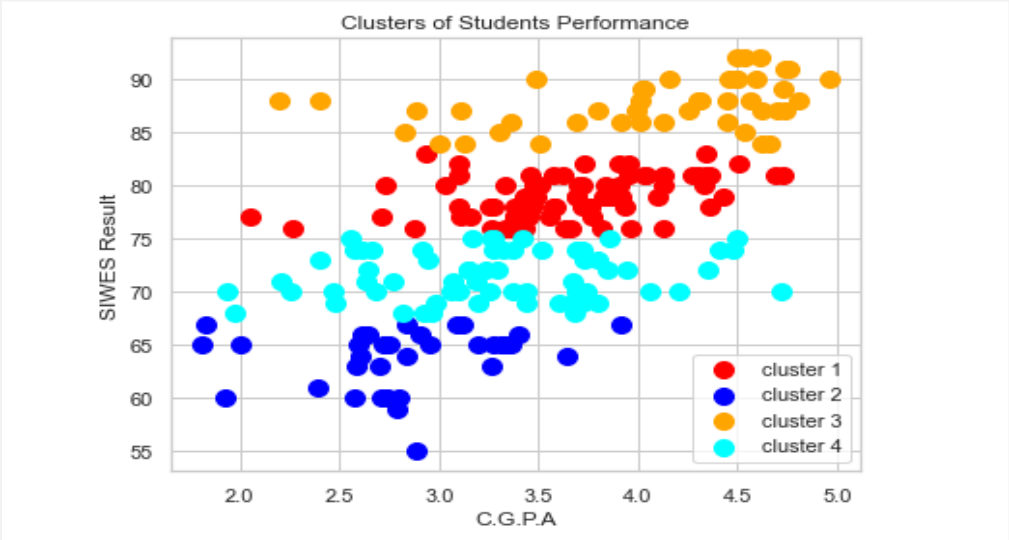

Figure 5: Four Clusters using 2 Variables (SIWES and CGPA)

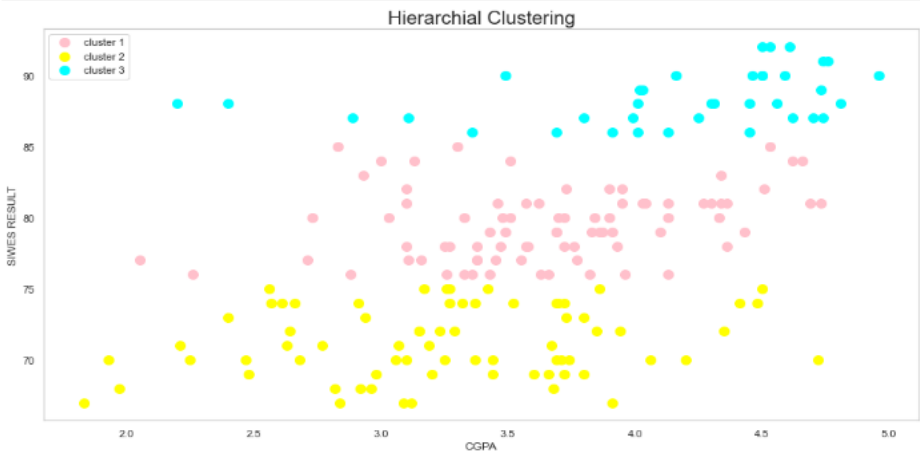

Figure 6: Hierarchy of 3 Clusters using Two Variables

Hierarchical clustering showed the hierarchy of data, and hence this hierarchy can be realized using a dendrogram, in which was visualized with 3 clusters. It is equally possible to have it clustered into 2 and 4 clusters just like with the K-means. Figure 6 is the hierarchical clustering of two variables (SIWES and CGPA) with 3 clusters. It indicates that students who scored 85% and above in SIWES and a good CGPA, have high possibility of gaining employment after graduation.

ii. Machine learning Classifiers

Though the dataset (218) used is small in size, 174 were trained and 44 tested using 3 machine learning classifications, to train, test, and build the models. RF, LR and DT were used in predicting the performance which shows best accuracy for all the algorithms. Considering (i) Our data set is not much (ii) We use algorithms when creating labels (i.e. K-means clustering) because the algorithm studies the data correctly and labels them and found it to be perfect. That is why; Figure 16 shows the confusion matrix of the model.

Table 3: Result of Performance from the Models Developed.

Algorithm	Accuracy	F1 Score	Precision	Recall
Random Forest	98%	0.99	0.90	0.96
Logistic Regression	94%	0.96	0.88	0.84
Decision Tree	97%	0.96	0.99	0.92

Using the best performing classifier (RF), we consider 44 instances (Testing set) and plot the confusion matrix in fig 7, the model classified 24 students that will not get employment and 20 students who will get employment after graduation correctly represented in the blue boxes. The white boxes with 0's signify Type I and Type II error and the model perform well. The upper right box signifies that 0 are unemployed and the model predicted 0 students will be employed. While the lower left box signifies 0 are going to be employed and the classifier predict 0 will be unemployed. Table 2 shows the performance results of the models.

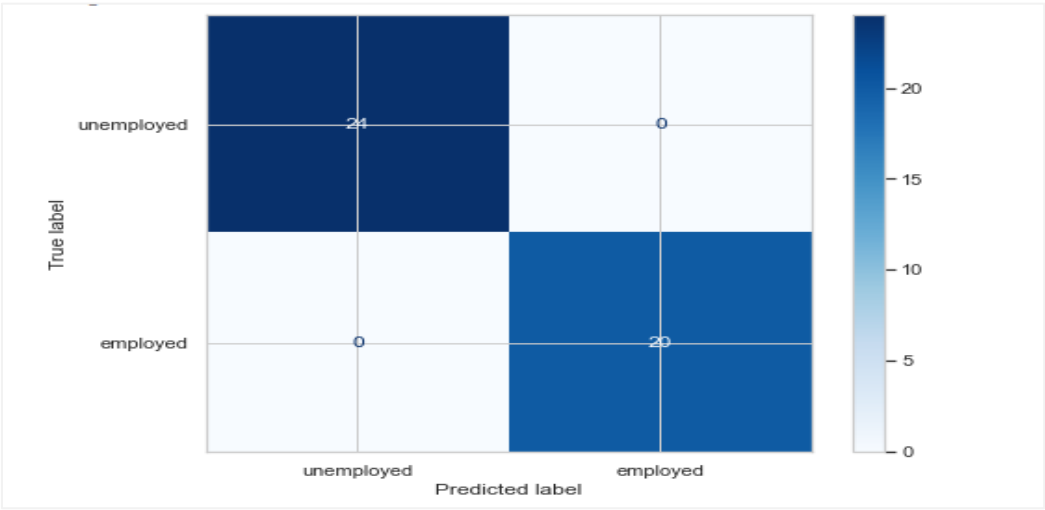

Figure 7 Confusion Matrix of the best performing model (Random Forest)

Using the Random Forest Classifier, SIWES result, CGPA are the most important in predicting students' employability. We plot the feature importance in terms of mean decreases in impunity as presented in figure 8.

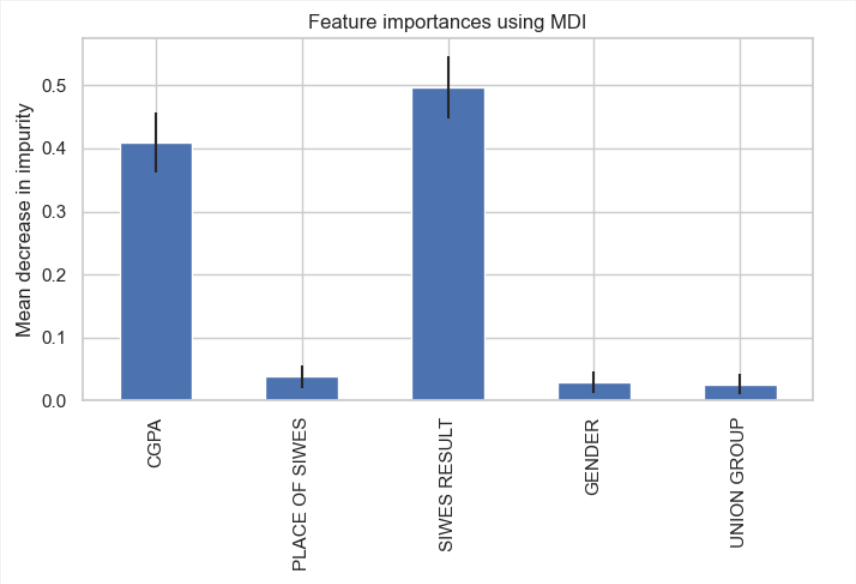

Figure 8: Random Forest Feature importance

Discussion

Our experiment has shown the effect of labeling the data using k-means clustering, which is due to the lack of label data in our disposure. The machine learning classifiers deployed in this study can predict students' employability with little or no error.

From the result of our work we can answer the research questions as follows;

In the literature review we have seen the extend of using various employability signals in predicting student chance of getting employment, in our study the attributes are limited to few like CGPA, SIWES result, union group, gender, year of graduation and so on.

Random Forest classifier performs better than the decision tree and logistic regression in terms accuracy and f1-score.

The k-means clustering algorithms depicted that SIWES result is the most important in grouping students based on their academic and extracurricular activities attributes while Random Forest Classifier ranked SIWES result, CGPA and place of SIWES as the most important attributes in predicting students' employability.

CONCLUSION

The study predicted employability of undergraduate students before graduation, while using academic performance and internships as preferences. Investigating the effectiveness of different machine learning techniques for predicting students' performance as an indicator of employment before graduation was carried out using unsupervised predicting algorithm for student likelihood employment possibility/opportunity after graduation was done perfectly. Comparison of machine learning model using four (4) metrics, precision, recall, F1 score accuracy, was done with RF recording 98% accuracy, 0.96 Recall and better F1 score of 0.99. Decision Tree classifier recording the best precision among the machine learning classifiers with 0.96, while Logistic Regression has the least accuracy of 94%. Further works can emphasis on collecting more data to make the predictive model unbiased and experiment different machine learning models.

REFRENCES

Esquivel, J. A., & Esquivel, J. A. (2020). Using a Binary Classification Model to Predict the Likelihood of Enrolment to the Undergraduate Program of a Philippine University. *International Journal of Computer Trends and Technology*, *68*(5), 6–10. **[Crossref]**

Guo, T., Xia, F., Zhen, S., Bai, X., Zhang, D., Liu, Z., & Tang, J. (2020). Graduate employment prediction with bias. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, 670–677. **[Crossref]**

Hugo, L. S. (2018). *A Comparison of Machine Learning Models Predicting Student Employment*. http://rave.ohiolink.edu/etdc/view?acc_num=ohiou1544127100472053

Kumar, M. S., & Babu, G. P. (2019). Comparative Study of Various Supervised Machine Learning Algorithms for an Early Effective Prediction of the Employability of Students. *Journal of Engineering Sciences*, *10*(10), 240–251.

Mezhoudi, N., Alghamdi, R., Aljunaid, R., Krichna, G., &

Düştegör, D. (2021). Employability prediction: a survey of current approaches, research challenges and applications. *Journal of Ambient Intelligence and Humanized Computing*. **[Crossref]**

Mishra, T., Kumar, D., & Gupta, S. (2017). Students' Performance and Employability Prediction through Data Mining: A Survey. *Indian Journal of Science and Technology*, *10*(24), 1–6. **[Crossref]**

Oladokun, V. O., Ph, D., Adebanjo, A. T., Sc, B., & Ph, D. (2008). *Predicting Students ' Academic Performance using Artificial Neural Network : A Case Study of an Engineering Course* . *9*(1), 72–79.

Palacio-Niño, J.-O., & Berzal, F. (2019). *Evaluation Metrics for Unsupervised Learning Algorithms*.

Vinutha, K., & Yogisha, H. K. (2021). Prediction of employability of engineering graduates using machine learning techniques. *Proceedings of the 2021 8th International Conference on Computing for Sustainable Global Development, INDIACom 2021*, 742–745. **[Crossref]**

Wanjau, S. K., & Muketha, G. M. (2018). Improving Student Enrollment Prediction Using Ensemble Classifiers. *International Journal of Computer Applications Technology and Research*, *07*(03), 122–128. **[Crossref]**