Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 17, no. 4, April 2021, pp. 371-378
DOI: 10.23940/ijpe.21.04.p5.371378

Performance Analysis of the Impact of Technical Skills on Employability

Manjushree D. Laddha[a,*], Varsha T. Lokare[b], Arvind W. Kiwelekar[a], and Laxman D. Netak[a]

[a]*Dr. Babasaheb Ambedkar Technological University, Lonere, 402103, India*
[b]*Rajarambapu Institute of Technology, Uran Islampur, 415414, India*

Abstract

The competency to predict student success in a course or program generates opportunities to enhance educational outcomes to improve graduate employment. With effective performance prediction techniques, teachers can appropriate resources and instruction more precisely. Research in this area aspires to recognize features that can be used to make predictions with the help of machine learning techniques that can refine predictions and quantify aspects of student performance on employability. Moreover, research in predicting student performance on employability strives to discover interrelated features and to connect the underlying reasons why definite features work better than others. This study is to build the Technical Skills Based Employability Prediction Model (TSBEPM) using machine learning techniques. The technical skills are the scores of the students in various programming courses. The experimental work is based on the predictions obtained by various machine learning classifiers, namely Support Vector Machine, Naive Bayes, Logistic Regression, Decision Tree, Random Forest, AdaBoost, and Artificial Neural Network. To confirm all models used, the experiments were carried out using real data collected from the graduate students at the University. With the help of performance measuring parameters, different models are formulated to be used for predicting whether a student is placed or not. Random Forest gives a maximum accuracy of 70% and F1-Score of 0.85. The model is formulated to be used for predicting whether a student is placed or not.

Keywords: machine learning techniques; classification model; technical skills; performance

1. Introduction

Baccalaureate employability is an international issue due to the growing number of IT degree holders generated by graduate institutes each year. The country is facing a big problem regarding unemployment of their Institute graduates, which may indicate that they are deficient in the proper competencies needed by employers. For this reason, a comprehensive study should be taken into consideration to recognize the important factors underlying graduate employability.

In the education-related scholarly literature, work on determining factors that contribute to employability has existed for at least for a century. With many educational institutes not producing significant employment chances to, a generation of youthful high-yielding graduates will express an unresolved future unless something special is done to increase their employability. To improve the graduates' chances of getting decent jobs that match their education and training, instructors need to equip their students with the important competencies to enter into the field.

Machine learning is a tool used to describe the analysis and search for desirable connections, such as patterns, correlation, and changes among variables in a dataset. Many machine learning methods can be used to find important and relevant knowledge from huge education data. Machine learning has several tasks, such as classification, prediction, and clustering. However, classification is one of the important methods in machine learning. Models created by classification techniques are accurately used to predict future data trends. Many algorithms are available for data classification, including Decision Tree, Naive Bayes, Logistic Regression, Artificial Neural Network classifiers, and so on.

Furthermore, the Random Forest is one of the most frequently used methods. It creates multiple decision trees from the data and merges them to get a more accurate and stable prediction. Random Forest algorithm can be used for both classification and regression types of problem.

* Corresponding author.
E-mail address: mdladdha@dbatu.ac.in

This research aims to predict student employability by using a technical skill-based employability prediction model to predict whether the student has been employed or not. The Technical Skills Based Employability Prediction System (TSBEPS) is proposed in this paper. The impact of technical skills on employability has been analyzed in this research work. Students' performance in various technical courses, namely C Programming, Data Structures and Algorithms, Mobile App Development, Machine Learning, and Web Technology are considered for study purposes.

The data is collected from the Department of Computer Engineering at Dr. Babasaheb Ambedkar Technological University, Lonere – Raigad, in the academic year 2018-2019 and 2019-2020. To select important features from the collected dataset, two feature selection methods have been applied, namely Pearson and Kendall Correlation. It has been observed that the scores in the C Programming, Data Structures and Algorithms, and Machine Learning courses show that there is no strong relationship between them, but it does affect the employability of the students. However, there is a negative correlation between the scores in Mobile App Development and Web Technology courses.

A total of 133 students' data has been considered for analysis purposes along with their performances in five courses. Seven machine learning models, namely Decision Tree, Random Forest, Support Vector Machine, Naive Bayes, Artificial Neural Network, AdaBoost, and Logistic Regression are applied for training and testing the proposed system. It has been observed that the Random Forest Employability prediction model gives a maximum accuracy of 70% and F1 score of 0.85 in predicting correct employment and unemployment.

2. Related Work

Several researchers used machine learning techniques in the education field for predicting certain behaviors of students, including performance, dropout, extracting rules, and modifying systems. Some have performed a systematic literature survey [1] in the area of student performance prediction and student placement. The research problem of students' performance prediction can be analyzed through different angles. Even student concentration levels [2] can be measured.

Some predictions of the employability of students can be carried out by the combination of the different methods like the Bayesian method and the Tree method techniques [3] in the education field. Tools like [4] WEKA, Rapid Miner, and R programming were used by considering the student profile data. Some researchers predict student performance and placement by applying Decision Tree algorithms [5], which were implemented by using the Rapid Miner tool. Researchers predict the employability of students by using the Naive Bayes Algorithm with different data mining tools [6]. To find better tools that provide the highest prediction accuracy on student placement data, various tools were analyzed.

As huge data is generated when using dashboard data, the software product line is used to analyze University employment [7]. Software engineering concepts like domain engineering, code generation, interoperability, personalization, and reusability were considered.

The different employability skills like creativity skills, conversation skills, analytical skills, decision -making skills, proactive skills, teamwork skills, adaptability skills, are analyzed and based on that, competency-based training [8] is conducted for the employability of different programs of engineering.

In [9] the researcher proposed his algorithm and compared it with three other classification algorithms: Decision tree, Naive Bayes, and Neural network.

This research aims to predict a student's employability by using a technical skill-based employability prediction model to predict whether the student has been employed or not. The main contribution of this research is to find the performance measuring parameters like accuracy and F1 Score of these seven algorithms from commonly used machine learning techniques in the education environment.

3. Proposed Technical Skills Based Employability Prediction Model (TSBEPM)

The machine learning model has been used to find out the programming courses in the curriculum that impacts the placement of the students. Hence, the performance of the students in various programming courses are the input to the machine learning model. The proposed model helps to predict whether a student will get a place in any company or not.

As shown in Figure 1, the proposed model will give predictions based on the students’ performance in various technical courses. The details of the actual working procedure of the system are given below:

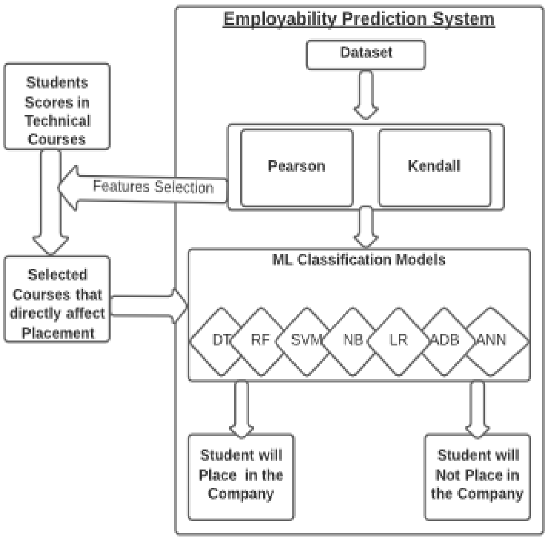
Employability Prediction System
Dataset
Students Scores in Technical Courses
Features Selection
Pearson
Kendall
Selected Courses that directly affect Placement
ML Classification Models
DT
RF
SVM
NB
LR
ADB
ANN
Student will Place in the Company
Student will Not Place in the Company

Figure 1. Proposed employability prediction model (EPM)

4. Dataset

Data is collected from the Dr. Babasaheb Ambedkar Technological University, Lonere - Raigad in the academic year 2019-2020. A total of 133 students' data regarding marks in various courses, namely C programming, Data Structures and Algorithms, Web Technology, Machine Learning, and Mobile Application Development are considered for experimental purposes. A sample dataset is shown in Table 1. Also, the placement status of the Department of Computer Engineering has been used for the analysis.

Table 1. Sample dataset

Roll No.	Marks in Computer Programming	Marks in Data Structures	Marks in Web Technology	Marks in Mobile App Development	Marks in Machine Learning	Placement
101	78	98	77	65	66	Yes
102	56	60	62	63	55	No
--	--	--	--	--	--	--

5. Features Selection

It is necessary to find out which courses are directly affecting placement. Hence, two feature selection methods have been applied here, namely Pearson and Kendall. It is observed that out of five courses, only the scores of three courses directly affect employability.

The Pearson Feature Selection method is a filter method in which only a subset of the selected features is chosen. It provides a correlation matrix as shown in Figure 2. Performance in three courses, namely C programming, Data Structure and Algorithms, and Machine Learning affect the placement.

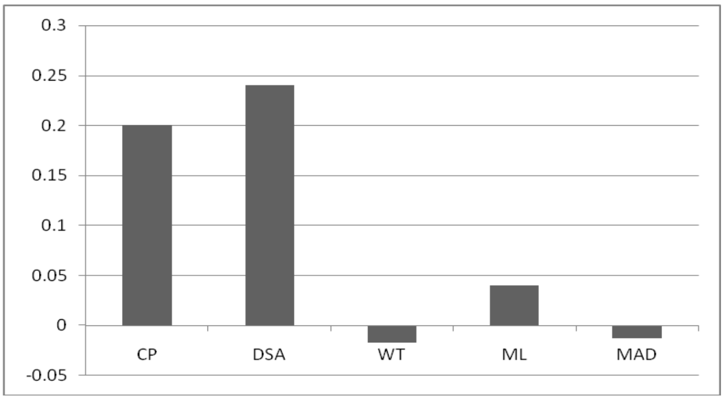

Figure 2. Pearson correlation method

The other two course scores, Web Technology and Mobile Application Development, do not at all affect the placement parameter as it shows a negative correlation. It shows the correlation value in the range of +1 to -1. As shown in Figure 3, the Kendalls method measures the ordinal relationship between the variables. This method worked on the normalization technique.

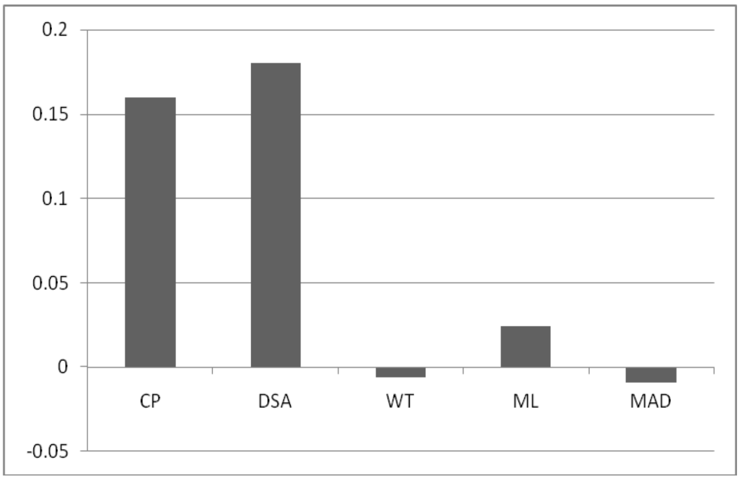

Figure 3. Kendall's correlation method

6. Machine Learning Classification Models

The problem is classified into two main classes: students will place or not place in a company. For this classification, the technical skills the students' scores have been considered. To predict the employability of the students, a total of seven ML classifiers have been considered: Decision Tree, Random Forest, Support Vector Machine, Naive Bayes, Artificial Neural Network(ANN), Adaptive Boost, and Logistic Regression.

6.1. Decision Tree

This algorithm can be applied for classification as well as regression purposes [10]. The Decision Tree classifier is based on a supervised learning approach. This model helped in the classification of the student's employability among hired or not hired. It uses a tree-like structure for this classification. To build a decision tree for this prediction, the Information Gain (IG) and Gini Index of each attribute (marks scored in different technical courses) have been calculated. The attribute with the highest IG score is considered as a root of the Decision Tree. Based on the IG score, nodes were further split and reach up to the desired prediction. Here, the outcome expected is the correct prediction of the student's employability. This paper mainly focused on analyzing the student's performance based on technical skills.

6.2. Random Forest

Similar to the Decision Tree algorithm, this algorithm can be applied for regression or classification problems [11]. As this ML classifier considered Multiple Decision Trees voting's for the prediction of the employability, the results are more promising than Decision Tree. Also, the Decision Tree algorithm mostly suffers the problem of over fitting, which is removed in the Random Forest algorithm. Also, the missing values issue is automatically solved in this type of classifier. Unlike Decision Tree, feature importance is calculated by the Random Forest model, which helps in understanding the most useful feature in the correct prediction of employability.

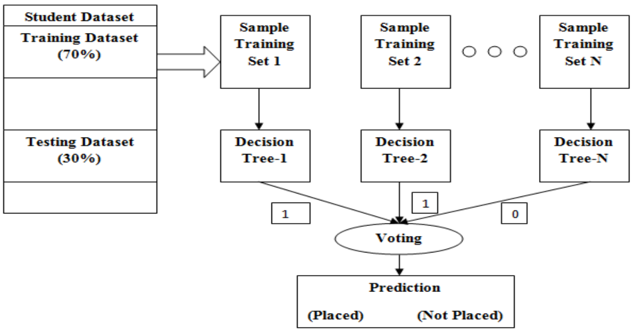

Figure 4. Random forest model for employment prediction

As shown in Figure 4, the opinion for all N Decision trees is considered and the prediction with maximum voting count will be the outcome of the Random Forest mode.

6.3. Support Vector Machine

This algorithm can be applied for regression or classification problems [12]. It performs well when we have a small dataset. From the scikit-learn, the model is implemented. Then, the SVM classifier is created by using a linear kernel on three features of the dataset to classify their class. The type of hyperplane used to separate the data is selected by kernel parameters. To fit the model, the gamma value is set to auto. It tries to exactly fit the training data set by a higher gamma value. Train the model for the training set and then predict the response of the testing set.

6.4. Naive Bayes

Naive Bayes is an extension of the Bayes Theorem [13]. The classifier understands that features have no relation with each other. The Naïve Bayes model is built with the help of scikit-learn by using the GaussianNB algorithm. This classifier is the combination of multiple algorithms and there is no correlation between the features. The training process is somewhat fast in this type of classification as probability measurement is the only task that needs to be calculated for the input features. Also, there is no need to fit the coefficient explicitly for the optimization purpose. This technique is mainly based on class and conditional probabilities. The basic assumption in this classifier is that all features equally contribute to the output, and there is no relation between the input features, i.e. in our case, it assumes that there is no correlation between the scores in various courses like C, Data Structure and Algorithms, etc. Also, scores in each course are equally contribute to employability. Hence, the results are not that promising. The class probability P(y) is as shown in Table 2:

Table 2. Class probability

Placement		P(Y)
Yes	90	90/133
No	43	43/133
Total	133	100%

6.5. Artificial Neural Network

Artificial Neural Network classifier is based on the human brain structure. Nodes are connected with the neurons in the brain [14]. While training the model and predicting the output, each node in the network gets input from the other nodes. As shown in Figure 5, the ANN model classifies the output into two classes:
Class1: Student will be placed
Class2: Student will not be placed

The performance in various courses like C programming, Data Structures, and machine learning is provided as input to the model. The ANN model uses the perceptrons to process the input and pass it to the hidden layer and finally reach the output layer.

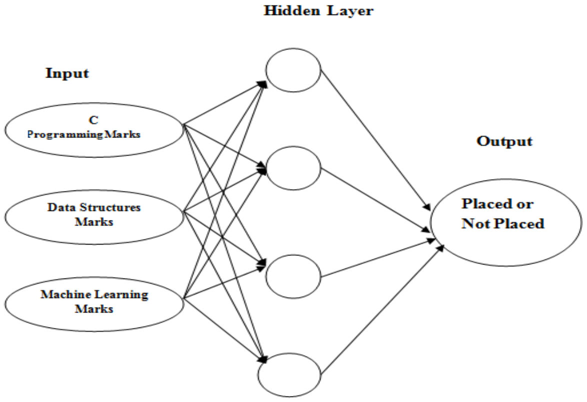

Figure 5. Artificial Neural Network Model for Employment Prediction

6.6. Adaptive Boost

Many weak classifiers like Decision Tree are combined into a single strong classifier [15]. It is also used for text classification problems. This classifier may tend to improve the performance as it works on the principle of assigning more weightages to the instances that are incorrectly predicted in the training phase. Also, it reduces the weight of the correctly predicted instances in further training.

6.7. Logistic Regression

This is generally used for binary classification. There are many applications of this classifier in data mining [16]. To predict the result in terms of employment and unemployment this classifier is used here. This model is based on the linear regression concept that measures the importance of every feature with the predicted output. The Logistic Regression algorithm is based on the phenomenon of the binary classification approach. There are two types present in this regression: one is locally weighted logistic regression and another is globally weighted regression. Generally, for complex problems, globally weighted regression is preferred. In this application, for the prediction of employment, weights are assigned locally and it is observed that the results are promising enough.

7. Model Evaluation and Validation

7.1. Experimental Work

As shown in Table 1, five courses were initially considered to measure the impact of technical skills on placement. Pearson and Kendall feature selection methods were used to find the actual affecting parameters on placement. It is observed that out of 5 courses, only 3 courses related to employability i.e., performance in C programming, Data Structures and Algorithms, and Machine Learning. Hence, for experimental purposes, only the scores in the above-mentioned courses are considered. A total 133 students' performances are considered here as input. 70% of the dataset is considered (93 entries) for training and 30% (40 entries) is used for testing purposes. Seven machine learning models are compared and analyzed: Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes, Logistic Regression, AdaBoost, and Artificial Neural Network (ANN).

7.2. Performance Measuring Parameters

To measure the performance of the machine learning model, a total of four measuring parameters are considered: Accuracy, Precision, Recall, and F1 Score.

i. Accuracy: The ratio of correct predictions to total predictions. To measure the percentage of correct predictions over total predictions, an accuracy metric is used.

$$\textit{Accuracy} = \textit{True Negative} + \textit{True Positive} / (\textit{True Positive} + \textit{False Positive} + \textit{False Negative} + \textit{True Negative}) \tag{1}$$

ii. Precision: The correct predictions that are positive by the following formula:

$$\textit{Precision} = \textit{True Positive} / (\textit{True Positive} + \textit{False Positive}) \tag{2}$$

iii. Recall: The number of positive class predictions made out of all positive examples in the dataset.

$$\textit{Recall} = \textit{True Positive} / (\textit{True Positive} + \textit{False Negative}) \tag{3}$$

iv. F1-Score: A single value based on precision and recall.

$$F1\ Score = 2 * (Precision * Recall) / (Precision + Recall) \tag{4}$$

7.3. Observations and Result Analysis

As shown in Table 3, Naive Bayes and Artificial Neural Network classifiers have the highest accuracy (75%) when compared to other machine learning models. Also, Random Forest and Logistic Regression showed an accuracy of (70%). Hence Naïve Bayes and Artificial Neural Network models are more suitable for predicting the employability of candidates based on technical skills. To calculate the complete performance of the model, other parameters are also important like precision, recall, and F1 score as shown in Table 3. In the case of predicting employability, Random Forest and Logistic Regression classifiers have shown better performance with precision (1). To measure true employability, out of all true employabilities in the dataset, the Random Forest classifier gives the best performance. The Random Forest machine learning classifier is based on the working principles of many Decision Trees; hence, it shows better results than other classifiers.

From Table 3, F1 score of Random Forest is (0.85), which is greater than Logistic Regression (0.14). Further accuracy of both classifiers is (70%). So, out of all classifiers, the Random forest F1 score (0.85) predicted correct employability.

Table 3. Performance measuring parameters based on Machine Learning (ML) models

ML Classifier	Unemployment			Employment			Accuracy
	Precision	Recall	F1 Score	Precision	Recall	F1 Score	
Decision Tree	0.72	0.67	0.69	0.4	0.46	0.43	57.50%
Random Forest	0.74	1	0.57	1	0.4	0.85	70%
Support Vector Machine	0.68	0.93	0.78	0.33	0.08	0.12	65%
Navie Bayes	0.74	0.96	0.84	0.8	0.31	0.44	75%
ANN	0.81	0.81	0.81	0.62	0.62	0.62	75%
AdaBoost	0.7	0.85	0.77	0.43	0.23	0.3	65%

Also, in the case of unemployment values, Random Forest and Logistic Regression provided the maximum recall (1). As shown in the Confusion Matrix Table 4 measuring parameter, the maximum correct unemployment rate (67.50%) is given by Logistic Regression, and an also equally good unemployment rate (62.50%) is given by Random Forest. As this model considered threshold-based decision boundaries to classify the output, it has shown better results than other classifiers. Also, in predicting wrong employment and unemployment, Logistic Regression is better than other classifiers as it predicted (0%) wrong unemployment and the lowest wrong employment (12%) as compared to other classifiers.

Table 4. Confusion matrix for correct prediction of employability

ML Model	Correct Unemployment	Wrong Unemployment	Correct Employment	Wrong Employment
Decision Tree	45%	22.50%	12.50%	20%
Random Forest	62.50%	10%	7.50%	20%
Support Vector Machine	62.50%	5%	2.50%	30%
Navie Bayes	65%	2.50%	10%	22.50%
ANN	55%	12.50%	20%	12.50%
AdaBoost	57.50%	10%	7.50%	25%
Logistic Regression	67.50%	0%	2.50%	12%

8. Conclusion

The Technical Skills Based Employability Prediction System (TSBEPS) is proposed in this paper. The impact of technical skills on employability has been analyzed in this research work. Students' scores in various technical courses were considered for study purposes. A total of seven machine learning models were constructed. Out of these, the Random Forest gave an accuracy of (70%) and F1 score of (0.85) in predicting correct employment and unemployment.

Predicting employability is helpful for students, teachers, as well as universities. For students, it is helpful for knowing what to improve to match the current requirements of the industry. For teachers, it is helpful in giving feedback to students. This model also helps universities make policy decisions to improve employability.

This work can be further extended by including scores of other technical courses and enhancing the total count of the dataset. Also, the deep learning approach can be applied for better performance of the model.

References

1. Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V.V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., and Liao, S.N. Predicting academic performance: a systematic literature review. In *Proceedings of the companion of the 23rd annual ACM conference on Innovation And Technology In Computer Science Education*, pp 175–199, 2018.
2. Lokare, V.T. and Netak, L.D. Concentration level prediction system for the students based on physiological measures using the EEG device. In *International Conference on Intelligent Human Computer Interaction*, Springer, Cham, pp. 24-33, November 2020.
3. Sapaat, M.A., Mustapha, A., Ahmad, J., Chamili, K., and Muhamad, R. A classification-based graduates employability model for tracer study by MOHE. In *Proceedings of the International Conference on Digital Information Processing and Communications*, Springer, pp 277–287, 2011.
4. Daud, A., Aljohani, N.R., Abbasi, R.A., Lytras, M.D., Abbas, F., and Alowibdi, J.S. Predicting student performance using advanced learning analytics. In *Proceedings of the 26th international conference on world wide web companion*, pp. 415-421, April 2017.
5. Jeevalatha, T., Ananthi, N., Kumar, D.S. Performance analysis of under-graduate students placement selection using decision tree algorithms. *International Journal of Computer Applications*, 108(15), 2014.
6. Mavani, U., Lobo, V.B., Pednekar, A., Pereira, N.C., Mishra, R., and Ansari, N. Naive bayes classification on student placement data: A comparative study of data mining tools. *Information and Communication Technology for Sustainable Development*, *Advances in Intelligent Systems and Computing*, Springer, 933, pp 363–372, 2020
7. Vázquez-Ingelmo, A., García-Peñalvo, F.J. and Therón, R. Domain engineering for generating dashboards to analyze employment and employability in the academic context. In *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality*, pp. 896-901, October 2018.
8. Boahin, P. and Hofman, A. A disciplinary perspective of competency-based training on the acquisition of employability skills. *Journal of Vocational Education and Training,* 65(3), pp. 385–401, 2013.
9. Ashok, M. and Apoorva, A. Data mining approach for predicting student and institution's placement percentage. In *Proceedings of the IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, pp 336–340, 2016.
10. Xiaochen, D. and Xue, H. Decision-Tree Classifier in Master Data Management System. In *Proceedings of the International Conference on Business Management and Electronic Information*, IEEE, 3, pp 756–759, 2011.
11. Thakar, P. and Mehta, A. Role of Secondary Attributes to Boost the Prediction Accuracy of Students Employability Via Data Mining. arXiv preprint arXiv:1708.02940, 2017.
12. Casuat, C.D. and Festijo, E.D. Predicting students' employability using machine learning approach. In *Proceedings of the IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, pp 1–5, 2019.
13. Umadevi, S. and Marseline, K.J. A survey on data mining classification algorithms. In *Proceedings of the IEEE International Conference on Signal Processing and Communication (ICSPC)*, pp 264–268, 2017.
14. Adewole, P. and Okewu, E. Artificial neural network-based learning analytics technique for employability and self-sustenance, 2018.
15. An, T.K and Kim, M.H. A new diverse AdaBoost classifier. In *Proceedings of the IEEE International conference on artificial intelligence and computational intelligence*, 1, pp 359–363, 2010.
16. Maalouf, M. Logistic regression in data analysis: an overview. *International Journal of Data Analysis Techniques and Strategies*, 3(3), 281–299, 2011.