

A prediction model based machine learning algorithms with feature selection approaches over imbalanced dataset

Alaa Khalaf Hamoud¹, Mohammed Baqr Mohammed Kamel^{2,3,4}, Alaa Sahl Gaafar⁵,
Ali Salah Alasady⁶, Aqeel Majeed Humadi⁷, Wid Akeel Awadh¹, Jasim Mohammed Dahr⁵

¹Department of Computer Information Systems, University of Basrah, Basrah, Iraq

²Department of Computer Algebra, Eotvos Lorand University, Budapest, Hungary

³Department of Computer Science, Hochschule Furtwangen University, Furtwangen im Schwarzwald, Germany

⁴Department of Computer Science, University of Kufa, Kufa, Iraq

⁵Department of Educational Planning, Directorate of Education in Basrah, Basrah, Iraq

⁶Department of Computer Science, University of Basrah, Basrah, Iraq

⁷Department of Software Engineering, College of Software Engineering, Islamic Azad University of Khorasgan, Isfahan, Iran

Article Info

Article history:

Received Jan 10, 2022

Revised Aug 3, 2022

Accepted Aug 30, 2022

Keywords:

Educational data mining

Feature selection

SMOTE filter

Students' performance

Supervised Algorithms

Unsupervised Algorithms

ABSTRACT

The educational sector faced many types of research in predicting student performance based on supervised and unsupervised machine learning algorithms. Most students' performance data are imbalanced, where the final classes are not equally represented. Besides the size of the dataset, this problem affects the model's prediction accuracy. In this paper, the Synthetic Minority Oversampling TEchnique (SMOTE) filter is applied to the dataset to find its effect on the model's accuracy. Four feature selection approaches are applied to find the most correlated attributes that affect the students' performance. The SMOTE filter is examined before and after applying feature selection approaches to measure the model's accuracy with supervised and unsupervised algorithms. Three supervised/unsupervised algorithms are examined based on feature selection approaches to predict the students' performance. The findings show that supervised algorithms (logistic model trees (LMT), simple logistic, and random forest) got high accuracy after applying SMOTE without feature selection. The prediction accuracies of unsupervised algorithms (Canopy, expectations maximization (EM), and farthest first) are enhanced after applying feature selection approaches and SMOTE filter.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Alaa Khalaf Hamoud

Department of Computer Information Systems

College of Computer Science and Information Technology, University of Basrah

Iraq

Email: alaa.hamoud@uobasrah.edu.iq

1. INTRODUCTION

Predicting students' performance in educational institutions is a major source of worry and interest for many researchers and governments around the world. It is important for educational institutions to monitor their students' performance and take appropriate action [1], [2]. University educators should evaluate the performance of their students to achieve desired goals and promote an environment of continuous improvement [3]. Educational institutions need to improve their ranking, based on many factors that are taken into account to evaluate their performance. Institutions are interested in providing quality education to improve students' performance. Therefore, the students' performance plays a major role in the rankings of higher education institutions [4].

The massive size of educational data and the numerous variables inside these data led to implementing different supervised/unsupervised machine learning approaches to find hidden patterns and implement the prediction process. The prediction process is used with dropout, students at risk, factors influencing students' performance, students' success, learners' performance, and overall academic institution performance [5]. The field of educational data mining (EDM) is one of the scientific disciplines that involve extracting hidden patterns, discovering new trends, and developing technologies and methods to analyze educational data [6]–[8]. EDM is a new methodology of research that utilizes data mining techniques in educational institutions to obtain insights into the problems students face in achieving outstanding performance. One purpose of EDM is to upskill the students and execute remedial actions to improve students' poor performance and thus to offer better educational performance. Applying supervised and unsupervised machine learning algorithms in education is continuously attracting researchers from the EDM domain [5], [9]. The topics of students' performance analysis and performance prediction are the most explored in the field of education system literature due to their valuable impact on the academic prediction outcome. The valuable knowledge outcome from using sophisticated algorithms may help the academic stakeholders (academic staff and students) [10]. Predicting students' at risk, or drop out in students' performance can help the academic staff in determining the factors that influence the performance. Different approaches are used with predicting students' performance such as [11]–[14], analyzing and predicting learners' performance [15]–[18], and academic institution performance [19]. Many approaches are used in prediction such as supervised machine learning approaches [20]–[24], unsupervised machine learning approaches [25]–[29] and semi-supervised machine learning approaches [30]–[33].

The problem of imbalanced class is raised in many fields and sectors of data mining. The class is said imbalanced if it contains small amount in one class compared with other classes. The minority class is considered as the positive class while the negative class is the majority class [34]–[41]. The imbalanced classification, related to educational data mining, faced many researches and many models are implemented to handle the imbalanced class in order to improve the accuracy [42]–[46]. Many filters are used to increase the number of minority classes such as synthetic minority oversampling technique (SMOTE) with the educational data. In this paper, many research questions will be investigated such as:

- Can Feature Selection (FS) impact the prediction accuracy of both of supervised/unsupervised algorithms?
- How SMOTE filter affect the supervised/unsupervised algorithms prediction accuracy?
- What is the effect of SMOTE filter on the slight imbalanced dataset?
- What are the optimal supervised/unsupervised algorithm in predicting students' performance before/after FS with/out applying SMOTE filter?

The dataset in the proposed model consists of 161 records from the college of computer science and information technology, university of Basrah, Iraq, which obtained based on a questionnaire that collects students answers to many questions related to their academic behavior and performance, sports activities, healthy food, their budgets, and information about their parents. After performing data preprocessing and removing the dirty data, the remaining records are 151. The goal class indicates whether the students are failed in one course (at least) or not. Many previous works are implemented on the same dataset to predict the students' performance based on supervised and unsupervised machine learning approaches. In contrast, the final works involve implementing FS approaches to find the most correlated features that affect the final class. In this work, the effect of SMOTE filter on the prediction accuracy of both of the top three supervised and unsupervised machine learning approaches will be investigated with/without FS approach.

The rest of the paper includes listing the related works, discussing and critiquing the literature review in section 2. Section 3 includes the proposed model explanation, the roadmap of implementing and evaluating the results. The final section includes the concluded points from this model and the future works.

2. LITERATURE REVIEW

Ashraf *et al.* [47] proposed a model based on several ensemble techniques. The researchers examined boosting to predict the students' performance. They selected ensemble approaches as prediction and classification approaches due to their significance, hence they adopted boosting technique to get highly accurate results. They adopted two filters (SMOTE, and spread subsampling) with four base classifiers (Naïve Bayes, KNN, J48, and random tree). They found that both filtering and ensemble approaches had improvement in predicting g students' performance. As a results, two prediction models had been propounded based on the conducted improvement.

Desiani *et al.* [48] are proposed a model based on a dataset of Universitas Sriwijaya with 2,934 records. The researchers tried to identify and solve the minority class labels (tightest 27%, and very tight 38.6%). The data is cleansed by removing the missing values and SMOTE filter is applied on the data to handle imbalanced data. The evaluation process of the model is performed against three machine learning

algorithms (C4.5, artificial neural network (ANN), and K-nearest neighbor (KNN)). The results showed that there is a significant improvement in accuracy, precision, and recall of minority classes.

Hassan *et al.* [41] proposed a model based on 4,413 row data of two datasets: e-learning and students' information of the first semester of academic year 2017-2018 in the Malaysia university, faculty of engineering. Three different sampling filters are utilized: oversampling, undersampling, and hybrid techniques. Five ensemble classifiers are examined against seven sampling techniques. The results shows that the Random Oversampling (ROS) hybrid technique with AdaBoost outperformed the other techniques. The study shows that SMOTEEN with ensemble classifiers got the high accuracy for the prediction model. Next, Utari *et al.* [49] are analyzed 2,492 raw educational data from 2008-2012 based on data mining techniques. Classification approach is used to predict drop-outs for undergraduate students in the imbalanced data. SMOTE filter is implemented to handle imbalanced data and random forest algorithm is utilized to predict drop-out students. The study shows that SMOTE filter with random forest can predict the drop-out students with 93.43% accuracy.

Many models are implemented on the educational dataset in this paper, started from [50], [51] where Bayesian and decision tree approaches are implemented to predict the students' performance. Khalaf *et al.* [52] association rules are implemented with feature selection to find the dataset's hidden patterns and the most correlated features. The unsupervised machine learning clustering approaches are implemented with PCA as a feature selection approach [28]. Four feature selection approaches are implemented to find the correlated features and then applied an artificial neural network approach to predict the students' performance. However, in some papers, feature selection approaches are used to find the most correlated feature to the final class, while in other papers, supervised/unsupervised approaches are implemented to predict performance. SMOTE is used in many papers without feature selection approaches. In this paper, SMOTE filter is implemented with/without feature selection approaches to find how it affects the model's accuracy after examining supervised and unsupervised approaches.

3. RESEARCH METHODS

The implementation process of the model goes through five steps: data collection, data preprocessing, feature selection, implementing algorithms after applying SMOTE filter with/without feature selection, and result evaluation, as shown in Figure 1. The questionnaire implementing process is considered part of the data collection step. It involves preparing the dataset for the next steps, such as removing incomplete answers, improving the dataset's quality, converting columns' domains, and deriving the last column (the goal class). The goal class (Grade) is derived based on (number of failed courses) question in the questionnaire where it takes the value (F abbreviation of Failed) if the number of failed courses greater than 0, and (P abbreviation of Pass) if the number of failed courses equal to 0. The dataset after data preprocessing goes through two directions, the first is applying FS and the data mining algorithms, and the second is applying smote then applying feature selection to find the effect of SMOTE filter on the dataset before/after the feature selection process.

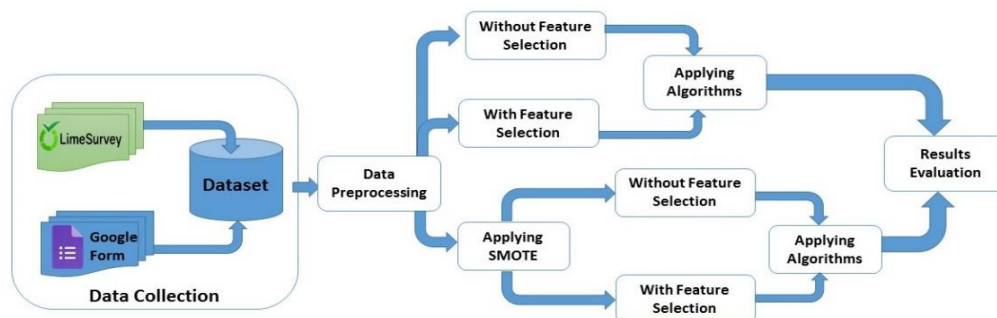


Figure 1. Model implementation diagram

3.1. Data collection

In the first step, the questionnaire is built using LimeSurvey open source web application and google form to collect students' answers. The two sources are merged in one Comma-Separated Value (CSV) file in order to prepare the data for the next step. The target students are the students in the College of Computer Science and Information Technology/University of Basrah. LimeSurvey is used for collecting the answers locally while Google form questionnaire is used for collecting the students' answers over the internet.

3.2. Data preprocessing

The total number of answers is 161 records with 61 questions. The sample is considered acceptable with a 10% margin of error [53]. The empty answers are removed, and the domains are converted to be evaluated and processed. Ten records were removed since they had missing values for many questions, and 151 answers were obtained. Weka 3.8.5 is used to visualize and implement data mining algorithms besides feature selection. The Cronbach's alpha value is present for the dataset to measure the dataset consistency and reliability, representing the overall internal uniformity among the columns. This value is used as a reliability indicator for the variables dependency for the study's dataset. The internal consistency is considered reliable when the value is 0.7 [54], [55]. In Table 1, the Cronbach's alpha value is 0.85 where it was measured for the 161 answers and 60 questions.

Table 1. Cronbach's alpha value

Number of respondents	Respondent Percentage	Number of Items	Cronbach's alpha
161	100%	60	0.85

3.3. Feature selection

The model requires two types of datasets, the first one with all features and the second one with selected features to measure the effect of feature selection on the result accuracy. Four feature selection algorithms (support vector machine (SVM), principal component analysis (PCA), info gain, and correlation) have been applied and the result ranks values of the features were observed as shown in the sample Table 2. The average value (AVG) of each question (QN) is measured related to the final class (Grade).

Table 2. Correlation according to feature selection algorithms (Sample) [30]

Rank	PCA	Info Gain		SVM		Correlation	
	QN	AVG	QN	AVG	QN	AVG	QN
1	61	0.173	15	57.5	26	0.451	15
2	19	0.062	18	57.2	30	0.315	26
3	20	0.055	61	54.7	10	0.233	47
4	18	0.013	8	53.5	58	0.237	16
5	22	0.002	6	53.4	18	0.236	61
6	17	0.002	4	52.3	33	0.235	18
7	21	0.002	5	49.8	41	0.211	48
8	23	0.001	1	49.4	15	0.211	52
9	30	0.051	26	46.5	49	0.186	10
10	27	0	7	42.9	16	0.168	14

The results are listed in an ascending order based on rank where each algorithm result differs from others. The average rank of each question is measured based on cross-validation. The first ranked question according to PCA is QN 61, where QN 15 is the first according to the Info Gain algorithm with AVG 0.173. QN 26 is the first according to the SVM algorithm with AVG 57.5, while QN 15 is the first according to Correlation with AVG 0.451. The other questions have been ranked using the same approach. The top 30 questions are selected based on the top average ranking of the correlation. For example, QN 1 will have an average rank based on dividing the summation of all feature selection algorithms rankings by four. The result average ranks of all questions are conducted to get the top 30 questions.

3.4. SMOTE filter

EDM techniques extract and discover new understanding and relationships from studying and analyzing the data that has been generated from various educational institutions such as schools and universities. Due to various reasons, this data might be available in limited quantity. One problem that can be caused by having limited quantity of educational data is the imbalanced dataset, in which the number of data points in one class (e.g. passed students) is considerably higher than number of data points in the other class (e.g. failed students). During analyzing such imbalanced dataset, most techniques will deal with the majority class (i.e. the passed students class) and treat the minority class (i.e. the failed students) as noise that might be ignored. If this issue is not addressed beforehand, it can negatively affect the results and outcomes of the mining. To address this issue, resampling (undersampling or oversampling) can be used to readjust the distribution of imbalanced dataset that results in improving the training power and increasing the accuracy of the prediction [56], [57]. Generally, oversampling is preferred in the case of precision, as the undersampling might remove some parts of the dataset that includes important information which possibly lead to model

overfitting [58]. On the other hand, the best oversampling techniques are those that instead of replicating minority class data, create new minority class data.

Synthetic minority over-sampling technique (SMOTE) [59] as a popular oversampling technique, generates more samples of the minority class to allow the classifier to be able to learn from. To avoid overfitting issue while expanding the minority class, SMOTE operates within the feature space of the dataset. For a selected minority class sample s , the k nearest neighbors $\{j_1, j_2, \dots, j_{k-1}, j_k\}$ of the sample s are found. Among the k nearest neighbors, SMOTE chooses a random sample j_r among those samples and computes the dst value as the difference in distance between samples s and j_r . It then multiplies the dst value by a random value r between 0 and 1 as shown in (1) to generate a new sample n in minority class.

$$n = s + (r \times (j_r - s)) \quad (1)$$

As it is clear from (1) the newly generated sample in the minority class lays randomly on the line that connects two already existed samples s and j_r . Figure 2 illustrates the sample generation technique using SMOTE.

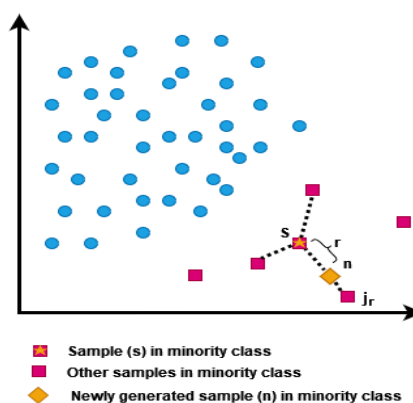


Figure 2. SMOTE oversampling example with $k=3$

3.5. Model Implementation

Weka 3.8.5 tool is used to implement and test the supervised/unsupervised machine learning algorithms. Weka tool is also used to apply SMOTE filter to the imbalanced dataset. The first round involves implementing the top three supervised machine learning algorithms logistic model trees (LMT), and random forest, and simple logistic. The top three unsupervised machine learning algorithms (Canopy, EM, and farthest first) are implemented also on the same dataset with all features. The second round involves implementing the same supervised/unsupervised machine learning algorithms with the dataset after selecting the top 30 ranked attributes. Several algorithms are used for classification such as decision tree algorithms, where these algorithms build a tree structure similar to a flowchart. The non-leaf nodes in the tree are a test on a specific variable, while the branches of the tree refer to the test results. The class label prediction refers to the external terminal node. The best variable to perform the division into classes is chosen to build the tree by the used algorithm [60]. Several algorithms have been developed to deal with numeric and categorical features in several sectors such as C4.5, ID3, RepTree, Hoeffding Tree, Random Forest, LMT, Random Tree, decision stump [50]. J48 is a decision tree that is an extension of ID3 and contains many features such as decision trees pruning, derivation of rules, continuous attribute value ranges and accounting for missing values [61]. A Bayesian network is a structure that indicates the conditional dependencies among the variables of the domain. Also, it is used to graphically show the potential underlying relationships among the variables of the domain. It is comprised of probability tables and directed acyclic graphs. The network node indicates the variables of the domain and an arc between two nodes represents having a basic relationship of subordination among the two nodes. Logistic regression is one of the statistical techniques which utilize for overcoming the least square regression problems. The value will be approximate when the value of the function is exceeded, so the linear regression will transform the target value based on logistic regression [62], [63].

In the unsupervised machine learning section, the clustering field includes many algorithms (Canopy, EM, farthest first, filtered clusterer, make density based, and simple K-Means). The Canopy clustering algorithm works in two stages, the first stage involves dividing the data into canopies then the distance measurements are made only among points in the common canopy [64]. EM algorithm are applied

in a genetic field the phenotype (the observed data) represents the function of the genotype (unobserved data). Another field of the area is used to estimate the mixture distributions parameters. EM is used in many fields such as the clinical sector, social, economic studies that require finding the hidden patterns and unknown variables that affect the outcomes [65], [66]. On the other hand, the farthest first traversal (FFT) algorithm is a greedy and fast algorithm that reduces the number of maximum radius of cluster. The meta-clusterer filtered clusterer algorithm permits to apply the filter before clusterer is been learnt. The training data form the shape of the structure clusterer while the test data processed by using the filter without affecting the structure [67]. All these algorithms are implementing with the dataset with all features after applying SMOTE filter and observing the performance accuracy. The dataset after removing the uncorrelated features and applying SMOTE filter are faced implementing supervised/unsupervised algorithms.

3.6. Results evaluation

The final step includes results evaluation of the supervised/unsupervised against four performance criteria namely True Positive (TP) rate, False Positive (FP) rate, precision, and recall. The base confusion matrix resulting from calculating the accuracy of the results is used to find these factors. The confusion matrix as shown in Table 3 consists of four categories: TP that represents the positive class that are correctly classified as positive, FP which represents the negative class that incorrectly classified as positive, True Negatives (TN) which represents the negatives that correctly classified as negatives, and False Negatives (FN) which represents the positives that incorrectly classified as negatives. TP rate measures the positive cases that correctly classified as positives, while the FP rate measures the negative cases that incorrectly classified as positives. The precision measures the positive cases that are classified as positives while the recall measures the relevance between cases in the dataset which represent the same value as TP rate [68]. These performance criteria are measured after implementing the supervised/unsupervised algorithm with the dataset, WithOut Feature Selection (WOFS) and With Feature Selection (WFS).

Table 3. Confusion matrix

Real	Predicate	
	Positive class	Negative class
Positive class	True Positive (TP)	False Negative (FN)
Negative class	False Positive (FP)	True Negative (TN)

The performance criteria are observed before/after implementing algorithms with dataset with/without FS before/after SMOTE filter. Three performance criteria are observed (precision, recall, F-measure) since the dataset are not the same after oversampling, so it could be hard to depend on TP, and FP rate. Precision represents the total number of correctly predictive cases according to totally predictive positive cases, while Recall represents the ration of correctly predictive cases to all cases. F-Measure represents the harmony measurement between precision and recall. Table 4 shows that Random Forest e highest value in precision with dataset WFS with 0.797, and the recall for the same algorithm is the highest with 0.795 and F-Measure with 0.789 with feature selection.

Table 4. Performance results of supervised machine learning algorithms

Algorithm	Precision		Recall		F-Measure	
	WOFS	WFS	WOFS	WFS	WOFS	WFS
LMT	0.708	0.733	0.709	0.735	0.695	0.729
Random forest	0.707	0.797	0.709	0.795	0.697	0.789
Simple logistic	0.708	0.733	0.709	0.735	0.695	0.729

The effect of SMOTE filter can be clearly observed with the accuracy of the algorithms in Table 5. SMOTE filter oversamples the dataset, so the accuracy increased with/without feature selection. Random Forest scores the highest value in precision with 0.83 with dataset before removing the features, while the same algorithm with the dataset after removing uncorrelated features still scores the highest value with 0.816. According to recall, the same algorithm scores the highest values with 0.83 before removing features, while the same algorithm scores the highest value with 0.816 after removing the uncorrelated features. For F-measure, random forest again scores the highest values with 0.83 before removing features, while the same algorithm scores 0.816 as a highest value after removing features.

Table 5. Performance results of supervised machine learning algorithms after applying SMOTE

Algorithm	Precision		Recall		F-Measure	
	WOFS	WFS	WOFS	WFS	WOFS	WFS
LMT	0.753	0.734	0.755	0.736	0.752	0.733
Random forest	0.830	0.816	0.830	0.816	0.830	0.816
Simple logistic	0.754	0.741	0.755	0.741	0.754	0.735

To clarify the improvement in the results after applying algorithms with dataset with and without FS before/after SMOTE, Figure 3 shows the precision, recall, and F-Measure values for all supervised algorithms. Figure 3 showed that the precision of the algorithms (LMT, random forest, and simple logistic) with/without feature selection before and after applying SMOTE filter. The improvements in precision values for the algorithms are clearly observed where the effect of SMOTE filter improve the precision metric for all algorithms. FS also improves the precision metric compared with applying algorithm before FS and it is obviously clear that the SMOTE improve the precision metric after FS. Random Forest without FS and after applying SMOTE filter scores the highest precision value.

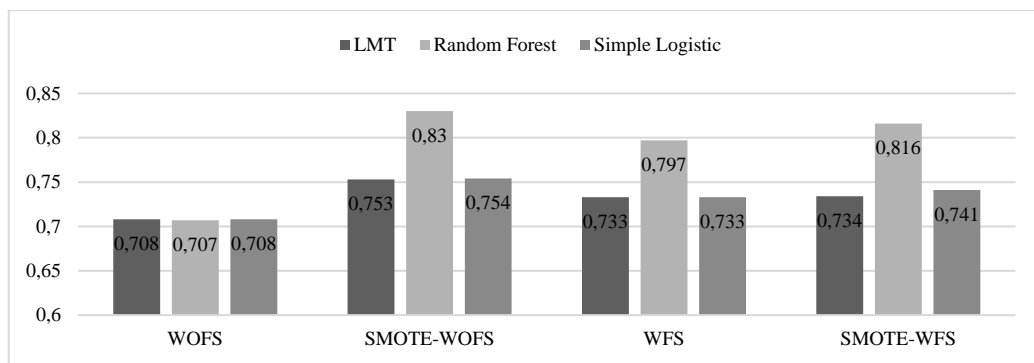


Figure 3. Precision of supervised algorithms

Figure 4 showed that the recall of the algorithms (LMT, random forest, and simple logistic) with/without FS before and after applying SMOTE filter. The improvements in recall metrics for the algorithms are clearly observed where the effect of SMOTE filter improves the recall metric for all algorithms before and after FS. FS also improves the recall metric compared with applying algorithm before FS and it is obviously clear that the SMOTE improve the recall metric after FS. Random forest without FS and after applying SMOTE filter scores the highest recall value.

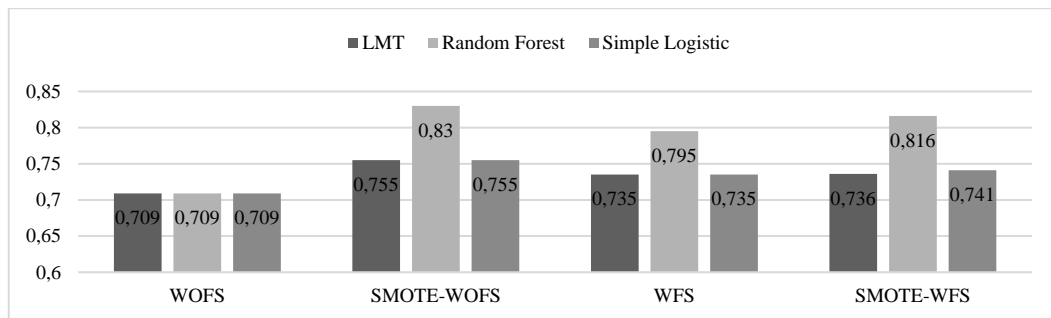


Figure 4. Recall of supervised algorithms

Figure 5 showed that the F-Measure of the same algorithms with/without FS before and after applying SMOTE filter. The improvements in recall metrics for the algorithms are clearly observed where the effect of SMOTE filter improves the F-Measure metric for all algorithms before and after FS. FS also improves the F-Measure metric compared with applying algorithm before FS and it is obviously clear that the SMOTE improve the F-Measure metric after FS. Random Forest without FS and after applying SMOTE filter scores the highest F-Measure value.

Next, Table 6 lists the performance results (precision, recall, and F-Measure) of the unsupervised machine learning algorithms in the field of clustering with/without feature selection before and after applying SMOTE filter. The clustering algorithms are the top three accurate algorithms (Canopy, EM, and farthest first). For the precision, Canopy algorithm outperformed the other algorithm with 0.635 average value with dataset WOFS, while the farthest first algorithm scored the lowest average value with 0.506 compared with other algorithms WOFS. After implementing feature selection, the average value of EM, and farthest first algorithms has been improved while Canopy precision is decreased. Regarding the recall with dataset WOFS, Canopy also scored the highest average value with 0.629 while farthest first algorithm scored the lowest value with 0.543. With dataset WFS, the recall average values of EM, and farthest first algorithms have been improved or remained same, while Canopy recall values is decreased. For the F-Measure, with dataset WOFS, Canopy scored the highest value with 0.618 while farthest first algorithm scored the lowest average value with 0.508. With dataset WFS, only EM, and farthest first algorithms have been improved from 0.574, and 0.508 to 0.577, and 0.531, and 0.568.

Table 7 shows the performance criteria with/without FS before/after applying SMOTE filter. SMOTE filter without FS enhances the precision metric for EM and farthest first while Canopy precision value decreased. Canopy, EM and farthest first algorithms precision values are enhanced with FS where the Canopy scored the highest values with 0.648. According to recall, again SMOTE filter without FS enhanced the recall metric for EM and farthest first while Canopy recall value is decreased. Later, Canopy, EM and farthest first algorithms recall values are enhanced with FS where the Canopy scored the highest values with 0.658. The same scenario for F-Measure without FS while Canopy scored the highest value with FS with 0.636.

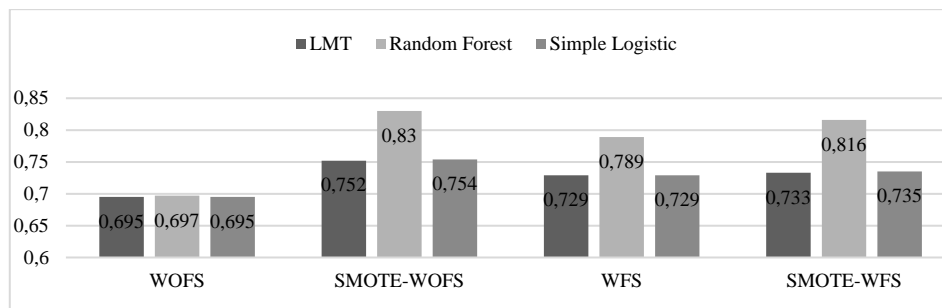


Figure 5. F-Measure of supervised algorithms

Table 6. Performance results of unsupervised machine learning algorithms

Algorithm	Precision		Recall		F-Measure	
	WOFS	WFS	WOFS	WFS	WOFS	WFS
Canopy	0.635	0.558	0.629	0.561	0.618	0.559
EM	0.578	0.584	0.571	0.574	0.574	0.577
FarthestFirst	0.506	0.527	0.543	0.543	0.508	0.531

Table 7. Performance results of unsupervised machine learning algorithms after applying SMOTE filter

Algorithm	Precision		Recall		F-Measure	
	WOFS	WFS	WOFS	WFS	WOFS	WFS
Canopy	0.595	0.648	0.618	0.658	0.596	0.636
EM	0.632	0.639	0.628	0.643	0.629	0.634
FarthestFirst	0.567	0.588	0.585	0.594	0.585	0.525

4. CONCLUSIONS

The total number of answers is 161 records with 61 questions. The sample is considered acceptable with a 10% margin of error. The empty answers are removed and the domains are converted in order to be evaluated and processed. Ten records were removed since they had missing values for many questions and a total of 151 answers were obtained. The final class is slight imbalance with 59.6% for fail and 40.4% for pass students. The model requires two types of datasets, the first one was with all features and the second one was with selected features to measure the effect of feature selection on the result accuracy. Four feature selection algorithms (SVM, PCA, Info gain, and correlation) have been applied and the result ranks values of the features were observed. The top three supervised machine learning algorithms (LMT, and random forest, and

simple logistic) besides the top three unsupervised machine learning algorithms (Canopy, EM, and farthest first) are implemented also on the same dataset with all features. All these algorithms are outperformed the other algorithms in the previous work. The supervised/unsupervised machine learning algorithms with the dataset after selecting the top 30 ranked attributes. For supervised algorithms, SMOTE filter oversample the dataset, so the accuracy increased with/without feature selection. Random Forest scores the highest value in precision with 0.83 with dataset before removing the features, while the same algorithm with the dataset after removing uncorrelated features still scores the highest value with 0.816. According to recall, the same algorithm scores the highest values with 0.83 before removing features, while the same algorithm scores the highest value with 0.816 after removing the uncorrelated features. For F-Measure, Random Forest again scores the highest values with 0.83 before removing features, while the same algorithm scores 0.816 as a highest values after removing features. For unsupervised algorithms, SMOTE filter without FS enhances the precision metric for EM and farthest first while Canopy precision value decreased. Canopy, EM and farthest first algorithms precision values are enhanced with FS where the Canopy scored the highest values with 0.648. According to recall, again SMOTE filter without FS enhanced the recall metric for EM and farthest first while Canopy recall value is decreased. Later, Canopy, EM and farthest first algorithms recall values are enhanced with FS where the Canopy scored the highest values with 0.658. The same scenario for F-Measure without FS while Canopy scored the highest value with FS with 0.636. In the future, different categories of filters such as oversampling, undersampling, and hybrid will be utilized to handle imbalanced data and find the most outperforming technique with and without feature selection approaches.

REFERENCES




- [1] R. Aggarwal and S. Pal, "Comparison of machine learning algorithms and ensemble technique for heart disease prediction," in *International Conference on Intelligent Systems Design and Applications*, 2021, pp. 1360–1370, doi: 10.1007/978-3-030-71187-0_126.
- [2] B. Siswoyo, Z. A. Abas, A. N. C. Pee, R. Komalasari, and N. Suyatna, "Ensemble machine learning algorithm optimization of bankruptcy prediction of bank," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 2, pp. 679–686, 2022, doi: 10.11591/ijai.v11.i2.pp679-686.
- [3] H. Talal and S. Saeed, "A study on adoption of data mining techniques to analyze academic performance," *ICIC Express Letters, Part B: Applications*, vol. 10, no. 8, pp. 681–687, 2019, doi: 10.24507/iceiclb.10.08.681.
- [4] W. F. W. Yaacob, S. A. M. Nasir, W. F. W. Yaacob, and N. M. Sobri, "Supervised data mining approach for predicting student performance," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 3, pp. 1584–1592, 2019, doi: 10.11591/ijeecs.v16.i3.pp1584-1592.
- [5] M. C. Sáiz-Manzanares, J. J. Rodríguez-Díez, J. F. Díez-Pastor, S. Rodríguez-Arribas, R. Marticorena-Sánchez, and Y. P. Ji, "Monitoring of student learning in learning management systems: An application of educational data mining techniques," *Applied Sciences (Switzerland)*, vol. 11, no. 6, p. 2677, 2021, doi: 10.3390/app11062677.
- [6] A. Pradeep, S. Das, and J. J. Kizhekkethottam, "Students dropout factor prediction using EDM techniques," in *Proceedings of the IEEE International Conference on Soft-Computing and Network Security, ICSNS 2015*, 2015, pp. 1–7, doi: 10.1109/ICSNS.2015.7292372.
- [7] S. Sivakumar, S. Venkataraman, and R. Selvaraj, "Predictive modeling of student dropout indicators in educational data mining using improved decision tree," *Indian Journal of Science and Technology*, vol. 9, no. 4, pp. 1–5, 2016, doi: 10.17485/ijst/2016/v9i4/87032.
- [8] B. Pérez, C. Castellanos, and D. Correal, "Predicting student drop-out rates using data mining techniques: A case study," in *Communications in Computer and Information Science*, 2018, vol. 833, pp. 111–125, doi: 10.1007/978-3-030-03023-0_10.
- [9] S. Tuaha, I. F. Siddiqui, and Q. Ali Arain, "Analyzing students' academic performance through educational data mining," *3C Tecnología_Glosas de innovación aplicadas a la pyme*, vol. 8, no. 1, pp. 402–421, 2019, doi: 10.17993/3ctecno.2019.specialissue2.402-421.
- [10] A. Khan and S. K. Ghosh, "Student performance analysis and prediction in classroom learning: A review of educational data mining studies," *Education and Information Technologies*, vol. 26, no. 1, pp. 205–240, 2021, doi: 10.1007/s10639-020-10230-3.
- [11] Y. S. Su and C. F. Lai, "Applying educational data mining to explore viewing behaviors and performance with flipped classrooms on the social media platform Facebook," *Frontiers in Psychology*, vol. 12, p. 653018, 2021, doi: 10.3389/fpsyg.2021.653018.
- [12] A. K. Hamoud, A. S. Hashim, and W. A. Awadh, "Predicting student performance in higher education institutions using decision tree analysis," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 2, p. 26, 2018, doi: 10.9781/ijimai.2018.02.004.
- [13] A. Khalaf, "Selection of best decision tree algorithm for prediction and classification of students' action," *American International Journal of Research in Science, Technology, Engineering & Mathematics*, vol. 16, no. October, p. 26, 2016.
- [14] J. A. Gómez-Pulido, A. Durán-Domínguez, and F. Pajuelo-Holguera, "Optimizing latent factors and collaborative filtering for students' performance prediction," *Applied Sciences (Switzerland)*, vol. 10, no. 16, p. 5601, 2020, doi: 10.3390/app10165601.
- [15] P. M. Moreno-Marcos, T. C. Pong, P. J. Muñoz-Merino, and C. D. Kloos, "Analysis of the Factors influencing learners' performance prediction with learning analytics," *IEEE Access*, vol. 8, pp. 5264–5282, 2020, doi: 10.1109/ACCESS.2019.2963503.
- [16] P. Shirsat, "Developing deep neural network for learner performance prediction in EKhool online learning platform," *Multimedia Research*, vol. 3, no. 4, pp. 24–31, 2020, doi: 10.46253/j.mr.v3i4.a3.
- [17] A. A. Mubarak, H. Cao, W. Zhang, and W. Zhang, "Visual analytics of video-clickstream data and prediction of learners' performance using deep learning models in MOOCs' courses," *Computer Applications in Engineering Education*, vol. 29, no. 4, pp. 710–732, 2021, doi: 10.1002/cae.22328.
- [18] R. Geetha, T. Padmavathy, and R. Anitha, "Prediction of the academic performance of slow learners using efficient machine learning algorithm," *Advances in Computational Intelligence*, vol. 1, no. 4, pp. 1–12, 2021, doi: 10.1007/s43674-021-00005-9.
- [19] A. F. Núñez-Naranjo, M. Ayala-Chauvin, and G. Riba-Sanmartí, "Prediction of University dropout using machine learning," in *International Conference on Information Technology & Systems*, 2021, pp. 396–406, doi: 10.1007/978-3-030-68285-9_38.

- [20] M. A. Muslim and Y. Dasril, "Company bankruptcy prediction framework based on the most influential features using XGBoost and stacking ensemble learning," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 6, pp. 5549–5557, 2021, doi: 10.11591/ijece.v11i6.pp5549-5557.
- [21] K. S. Nugroho, A. Y. Sukmadewa, A. Vidiyanto, and W. F. Mahmudy, "Effective predictive modelling for coronary artery diseases using support vector machine," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 1, pp. 345–355, 2022, doi: 10.11591/ijai.v11.i1.pp345-355.
- [22] O. Iatrellis, I. Savvas, P. Fitsilis, and V. C. Gerogiannis, "A two-phase machine learning approach for predicting student outcomes," *Education and Information Technologies*, vol. 26, no. 1, pp. 69–88, 2021, doi: 10.1007/s10639-020-10260-x.
- [23] S. D. Abdul Bujang, A. Selamat, and O. Krejcar, "A predictive analytics model for students grade prediction by supervised machine learning," in *IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1051, no. 1, p. 012005, doi: 10.1088/1757-899x/1051/1/012005.
- [24] M. Barramuño, C. Meza-Narváez, and G. Gálvez-García, "Prediction of student attrition risk using machine learning," *Journal of Applied Research in Higher Education*, vol. 14, no. 3, pp. 974–986, 2022, doi: 10.1108/JARHE-02-2021-0073.
- [25] N. Razali, S. Ismail, and A. Mustapha, "Machine learning approach for flood risks prediction," *IAES International Journal of Artificial Intelligence*, vol. 9, no. 1, pp. 73–80, 2020, doi: 10.11591/ijai.v9.i1.pp73-80.
- [26] M. Alenezi, M. Akour, and O. Al Qasem, "Harnessing deep learning algorithms to predict software refactoring," *Telkommika (Telecommunication Computing Electronics and Control)*, vol. 18, no. 6, pp. 2977–2982, 2020, doi: 10.12928/TELKOMNIKA.v18i6.16743.
- [27] S. Krishnan, P. Magalingam, and R. Ibrahim, "Hybrid deep learning model using recurrent neural network and gated recurrent unit for heart disease prediction," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 6, pp. 5467–5476, 2021, doi: 10.11591/ijece.v11i6.pp5467-5476.
- [28] A. K. Hamoud, "Classifying students' answers using clustering algorithms based on principle component analysis," *Journal of Theoretical and Applied Information Technology*, vol. 96, no. 7, pp. 1813–1825, 2018.
- [29] M. B. Janghel, et al., "Comparative study and analysis of students results using clustering techniques," *Information Technology in Industry*, vol. 9, no. 2, pp. 835–842, 2021, doi: 10.17762/itii.v9i2.421.
- [30] N. Uylaş, "Semi-Supervised Classification in educational data mining: students' performance case study," *International Journal of Computer Applications*, vol. 179, no. 26, pp. 13–17, 2018, doi: 10.5120/ijca2018916549.
- [31] V. Tam, E. Y. Lam, S. T. Fung, W. W. T. Fok, and A. H. K. Yuen, "Enhancing educational data mining techniques on online educational resources with a semi-supervised learning approach," in *Proceedings of 2015 IEEE International Conference on Teaching, Assessment and Learning for Engineering, TALE 2015*, 2016, pp. 203–206, doi: 10.1109/TALE.2015.7386044.
- [32] I. Hmiedi, H. Najadat, Z. Halloush, and I. Jalabneh, "Semi supervised prediction model in educational data mining," in *Proceedings - 2019 International Arab Conference on Information Technology, ACIT 2019*, 2019, pp. 27–31, doi: 10.1109/ACIT47987.2019.8991048.
- [33] I. E. Livieris, K. Drakopoulou, V. T. Tampakas, T. A. Mikropoulos, and P. Pintelas, "Predicting secondary school students' performance utilizing a semi-supervised learning approach," *Journal of Educational Computing Research*, vol. 57, no. 2, pp. 448–470, 2019, doi: 10.1177/0735633117752614.
- [34] M. A. Febriantono, S. H. Pramono, Rahmadwati, and G. Naghdy, "Classification of multiclass imbalanced data using cost-sensitive decision tree c5.0," *IAES International Journal of Artificial Intelligence*, vol. 9, no. 1, pp. 65–72, 2020, doi: 10.11591/ijai.v9.i1.pp65-72.
- [35] S. Uyun and E. Sulistyowati, "Feature selection for multiple water quality status: Integrated bootstrapping and SMOTE approach in imbalance classes," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 4, pp. 4331–4339, 2020, doi: 10.11591/ijece.v10i4.pp4331-4339.
- [36] A. S. Desuky, A. H. Omar, and N. M. Mostafa, "Boosting with crossover for improving imbalanced medical datasets classification," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2733–2741, 2021, doi: 10.11591/eei.v10i5.3121.
- [37] S. M. J. Moghaddam and A. Noroozi, "A novel imbalanced data classification approach using both under and over sampling," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2789–2795, 2021, doi: 10.11591/eei.v10i5.2785.
- [38] R. V. Kulkarni, S. Revathy, and S. H. Patil, "Smart pools of data with ensembles for adaptive learning in dynamic data streams with class imbalance," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 1, pp. 310–318, 2022, doi: 10.11591/ijai.v11.i1.pp310-318.
- [39] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-December, pp. 5375–5384, doi: 10.1109/CVPR.2016.580.
- [40] Q. Zou, S. Xie, Z. Lin, M. Wu, and Y. Ju, "Finding the best classification threshold in imbalanced classification," *Big Data Research*, vol. 5, pp. 2–8, 2016, doi: 10.1016/j.bdr.2015.12.001.
- [41] H. Hassan, N. B. Ahmad, and S. Anuar, "Improved students' performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining," in *Journal of Physics: Conference Series*, 2020, vol. 1529, no. 5, p. 52041, doi: 10.1088/1742-6596/1529/5/052041.
- [42] E. Ongko and Hartono, "Hybrid approach redefinition-multi class with resampling and feature selection for multi-class imbalance with overlapping and noise," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 3, pp. 1718–1728, 2021, doi: 10.11591/eei.v10i3.3057.
- [43] N. Rachburee and W. Punlumjeak, "Oversampling technique in student performance classification from engineering course," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 4, pp. 3567–3574, 2021, doi: 10.11591/ijece.v11i4.pp3567-3574.
- [44] I. Khan, A. R. Ahmad, N. Jabeur, and M. N. Mahdi, "Minimizing Classification errors in imbalanced dataset using means of sampling," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2021, vol. 13051 LNCS, pp. 435–446, doi: 10.1007/978-3-030-90235-3_38.
- [45] R. M. Mathew and R. Gunasundari, "An experimental study on the effect of resampling techniques in multiclass imbalanced data in learning sector," *Design Engineering*, no. 8, pp. 16216–16231, 2021.
- [46] R. K. Tripathi, L. Raja, A. Kumar, P. Dadheech, A. Kumar, and M. N. Nachappa, "A cluster based classification for imbalanced data using SMOTE," in *IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1099, no. 1, p. 012080, doi: 10.1088/1757-899x/1099/1/012080.
- [47] M. Ashraf, M. Zaman, and M. Ahmed, "An intelligent prediction system for educational data mining based on ensemble and filtering approaches," *Procedia Computer Science*, vol. 167, pp. 1471–1483, 2020, doi: 10.1016/j.procs.2020.03.358.




- [48] A. Desiani, S. Yahdin, A. Kartikasari, and Irmeilyana, "Handling the imbalanced data with missing value elimination smote in the classification of the relevance education background with graduates employment," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 2, pp. 346–354, 2021, doi: 10.11591/ijai.v10.i2.pp346-354.
- [49] M. Utari, B. Warsito, and R. Kusumaningrum, "Implementation of data mining for drop-out prediction using random forest method," in *2020 8th International Conference on Information and Communication Technology, ICoICT 2020*, 2020, pp. 1–5, doi: 10.1109/ICoICT49345.2020.9166276.
- [50] A. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student performance prediction model based on supervised machine learning algorithms," in *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 928, no. 3, p. 32019, doi: 10.1088/1757-899X/928/3/032019.
- [51] A. Khalaf, A. Majeed, W. Akeel, and A. Salah, "Students' success prediction based on bayes algorithms," *International Journal of Computer Applications*, vol. 178, no. 7, pp. 6–12, Nov. 2017, doi: 10.5120/ijca2017915506.
- [52] A. S. Hashim, A. K. Hamoud, and W. A. Awadh, "Analyzing students' answers using association rule mining based on feature selection," *Journal of Southwest Jiaotong University*, vol. 53, no. 5, pp. 1–16, 2018.
- [53] G. D. Israel, "Determining sample size," 1992.
- [54] B. Carson, "The transformative power of action learning," 2009. [Online]. Available: <https://www.chieflearningofficer.com/2009/08/20/the-transformative-power-of-action-learning/>.
- [55] U. Sekaran and R. Bougie, *Research methods for business: A skill building approach*, vol. 26, no. 2. john wiley & sons, 1993.
- [56] A. Yun-chung, "The effect of oversampling and undersampling on classifying imbalanced text datasets," Citeseer, 2004.
- [57] J. A. Jupin, T. Sutikno, M. A. Ismail, M. S. Mohamad, S. Kasim, and D. Stiawan, "Review of the machine learning methods in the classification of phishing attack," *Bulletin of Electrical Engineering and Informatics*, vol. 8, no. 4, pp. 1545–1555, 2019, doi: 10.11591/eei.v8i4.1922.
- [58] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.
- [59] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
- [60] I. A. Najm, A. K. Hamoud, J. Lloret, and I. Bosch, "Machine learning prediction approach to enhance congestion control in 5G IoT environment," *Electronics (Switzerland)*, vol. 8, no. 6, 2019, doi: 10.3390/electronics8060607.
- [61] G. Kaur and A. Chhabra, "Improved J48 classification algorithm for the prediction of diabetes," *International Journal of Computer Applications*, vol. 98, no. 22, pp. 13–17, 2014, doi: 10.5120/17314-7433.
- [62] A. Cayci, S. Eibe, E. Menasalvas, and Y. Saygin, "Bayesian networks to predict data mining algorithm behavior in ubiquitous computing environments," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6904 LNAI, Springer, 2011, pp. 119–141.
- [63] A. DeMaris, "A Tutorial in Logistic Regression," *Journal of Marriage and the Family*, vol. 57, no. 4, p. 956, 1995, doi: 10.2307/353415.
- [64] A. Kumar, Y. S. Ingle, A. Pande, and P. Dhule, "Canopy clustering: a review on pre-clustering approach to K-means clustering," *Int. J. Innov. Adv. Comput. Sci.(IJIACS)*, vol. 3, no. 5, pp. 22–29, 2014.
- [65] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996.
- [66] P. Chandre, P. Mahalle, and G. Shinde, "Intrusion prevention system using convolutional neural network for wireless sensor network," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 2, pp. 504–515, 2022, doi: 10.11591/ijai.v11.i2.pp504-515.
- [67] D. B. Dasari, G. Edamadaka, C. S. Chowdary, and M. Sobhana, "Anomaly-based network intrusion detection with ensemble classifiers and meta-heuristic scale (ECMHS) in traffic flow streams," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 10, pp. 9241–9268, 2021, doi: 10.1007/s12652-020-02628-1.
- [68] S. S. Pangastuti, K. Fithriasari, N. Iriawan, and W. Suryaningtyas, "Data Mining Approach for Educational Decision Support," *EKSAKTA: Journal of Sciences and Data Analysis*, pp. 33–44, 2021, doi: 10.20885/eksakta.vol2.iss1.art5.

BIOGRAPHIES OF AUTHORS






Alaa Khalaf Hamoud    is Associate Professor at college of Computer Science & Information Technology, University of Basrah, Iraq. He received BSc degree from Computer Science Department, University of Basrah in 2008 with first ranking college student. He also received his MSc degree with specialization in clinical data warehousing from the same department with first ranking department student. He participated in (seven months) IT administration course in TU Berlin-Germany. He is a lecturer in Computer Information Systems, University of Basrah, Iraq. His scientific interests are data mining, data warehousing. He can be contacted at email: alaa.hamoud@uobasrah.edu.iq.






Mohammed Baqr Mohammed Kamel    obtained his Master in Computer Science from the University of Baghdad and IT Administration from Technical University Berlin. He got the Minister of higher educations and deputy of Prime ministers of Iraq awards for holding the highest GPA in undergraduate program. Currently he is a PhD researcher in Eotvos Lorand University (ELTE) and Furtwangen University (HFU), a member of the Institute of Data Science, Cloud Computing and IT-Security. Recently, he got the Gold award from EIT Health Innovation day of having the best innovative project which later has been presented in EIT WinnerDay in Paris. His main research interests are in the area of network security and applied cryptography and mainly focus on security and privacy in distributed environments. He can be contacted at email: mohammedb.kamel@uokufa.edu.iq.






Alaa Sahl Gaafar    is the manager of the Information and Communications Department in the Directorate of Education in Basra. He holds a Bachelor's degree in Computer Science from the University of Basra and a Master's degree in Information Systems from the Osmania University, so he has the scientific title of Assistant Lecturer. The areas of research are networks and artificial intelligence (machine learning) and it is also concerned with everything related to information technology. He gave many lectures in information technology, network security, programming, communication networks and databases at the University of Basra. He can be contacted at email: alaasy.2040@gmail.com.






Ali Salah Alasady    was born in Basrah, Iraq in 1985. He earned a bachelor's degree in Computer Science from Basrah University in 2007 and a master's degree in the Information Technology field from the Tenaga University, Malaysia in 2014. He is working as a lecturer in the Department of Computer Science, Computer Science and Information Technology College, Basrah University. He has sixteen papers in the field of Computer Science (Information Security and Data mining and Cloud Computing). He can be contacted at email: ali_s.hashim@uobasrah.edu.iq.






Aqeel Majeed Humadi    received the B.Eng. degree in software engineering from Imam Ja'afar Al-Sadiq university, Iraq, in 2011 and the M.S. degrees in computer science from university of Basrah, Iraq, in 2014. Currently, he is a PhD student in the department of software engineering, faculty of computer engineering, Islamic Azad university, Isfahan (Khorasgan) branch, Iran. He works as an assistant chief engineer in information technology department, Missan Oil Company, Iraq. Previously, he worked in database section for 10 years in the same company. His research interests include medical image classification, medical health diagnose, and image retrieval. He can be contacted at email: aqeelm16@gmail.com.



Wid Akeel Awadh    was born in Basrah, Iraq in 1984. She earned a bachelor's degree in Computer Science from Basrah University in 2006 and a master's degree in the same area from the same university in 2012. She is working as a lecturer in the Department of Computer Information Systems, Computer Science and Information Technology College, Basrah University. She has sixteen papers in the field of computer science (information security and data mining cloud computing). She can be contacted at email: wid.jawad@uobasrah.edu.iq.



Jasim Mohammed Dahr    is an employee in the Information and Communications Department of the Directorate of Education in Basra, Iraq. He completed his Master's degree in Information Technology from the University Utara, Malaysia. His research areas include advanced databases and systems analysis, data warehouse, and machine learning and its algorithms. He has given many lectures at the College of Computer Science and Information Technology, University of Basra, Iraq. Also, he works as a lecturer at the College of Fine Arts. His research interests include prediction using machine learning algorithms as well as requirement tracking and advanced database design. He can be contacted at email: lec.jasim.dahr@uobasrah.edu.iq or jmd20586@gmail.com.