Date of publication JUN-30, 2024, date of current version DEC-22, 2023.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.23.2.3534

Determination of the Best Feature Subset for Learner Migration in Limpopo

FRANS RAMPHELE[1], ZENGHUI WANG[2], ADEDAYO YUSUFF[2]
[1]Department, Department of Computer Science, University of South Africa, Florida, Johannesburg, South Africa,1709
[2]Department of Electrical Engineering, University of South Africa, Florida, Johannesburg, South Africa,1709

Corresponding author: Frans Ramphele (e-mail: letsukulo.ramphele@ gmail.com).

ABSTRACT The South African Education Management Information Systems (EMIS) hosts longitudinal data on school inventory, learners, and educators. One of the most prevailing and yet ignored phases in machine learning is Feature Selection (FS). Neglecting this phase can adversely impact the outcome of the machine-learning exercise. This study seeks to explore informative features from the EMIS system which can predict the possibility of learners prematurely transitioning to alternative learning spaces in the Limpopo education system. The Ravenstein migration theory was used to assemble the initial features which were then subjected to Boruta, RPART, Adaboost.M1, and J48 algorithms. The feature subsets generated by the FS algorithms were compared with filter-based statistical methods such as Spearman Correlation and Mutual Information to aid in the final selection of the best feature subset for the study. All machine learning FS methods performed well. Feature subset generated by Boruta was considered optimal due to relatively low importance score variance among the selected features compared to RPART, J48, and Adaboost.M1. It is believed that the low variance in the feature set will improve the model's stability and its ability to generalize with previously unseen data.

KEYWORDS Boruta; RPART; J48; Adaboost.M1; Mutual-Information; Spearman Correlation; Feature-Selection; Learner Migration

I. INTRODUCTION

HUMAN migration has been studied extensively in the past three decades to understand the patterns and associated causes. Work in this area has progressed and has been used to understand, among other things, complex social and economic systems to ensure the future survival of humanity [1]. The study of migration is not specific to humans, it is a widespread phenomenon among various taxa and encompasses various research areas in both humans and animals such as anthropology, sociology, economics, and ethology [1, 2]. What all these migrations have in common is the importance of maximizing the use of resources and different fitness advantages such as growth, reproduction and protection [2].In developing countries, this is a topic of central importance, since teaching and learning resources are meagre [3]. In addition, learners often migrate or displace from one school to another due to various factors and intervening obstacles. This raises concern especially on budget and resource allocation in the receiving schools, mainly because resources are generally allocated at the beginning of the year and do not follow learner movements throughout the academic year. Furthermore, the arrival of learners puts more pressure on receiving schools on issues related to gaps in curriculum coverage and related imperatives that need to be addressed to improve learner achievement. This is worse when learner migration is high, increasing pressure on education authorities to place learners and disrupting teaching and learning. The systematic review by [4] examined several studies exploring different factors that contribute to student migration in South African schools and ways to mitigate their negative effects. This review found that the migration of learners from one school to another is influenced by several factors, including legal frameworks in the education system, school management and leadership practices of schools, school efficiency, infrastructure, and socio-economic factors, among others. Conceptually, the work by [5] made similar observations. Botha and Neluvhola [4] acknowledged resource deficiencies in planning, management, and resource allocation caused by learner migration and recommended the development of policies that will enable school leaders to effectively deal with learner migration. Simelani [5] raised concerns about the reduction in human capital in rural schools, the increasing number of non-viable schools with multi-grade classes in primary schools, and the reduction in subject streams in secondary schools due to learner migration.

The study by [6] highlighted the limitations of the school legal framework in South Africa, which gives learners the right to be admitted to a school of their choice. Hettie [6] argued that the current learner admission policy created an ongoing exodus of learners moving from traditionally black schools to the former Model C schools (*former white schools post-apartheid education*). The study further argued that social class and the language of the school also play an important role in guiding learners in their school choices, with English being the most preferred language for learning and teaching [6]. There is a consensus among researchers [4, 6] that more orderly and governed schools where teachers are motivated, dedicated principals, organized parental involvement, school infrastructure, school socioeconomic status, and performance-are the most important drivers of learner migration.

The phenomenon of learner migration is not isolated to the South African education system. It spans across various countries. What is common in learner migration studies, is the growing need for access to quality education and to maximize life opportunities [7–9]. In recent years there has been a growing interest in researching the use of data mining techniques to solve problems in educational settings under the umbrella term Education Data Mining (EDM) [10]. One such study was conducted by [11] using cloud computing and data mining methods to assess the mental health of immigrant students moving to Chinese cities. To select the best features for the study, the data collected via the survey was first classified and fed into the feature selection function of the LTSM time-series neural network. Data from the selected features was further clustered using K-Means to reveal underlying patterns to help answer the research question. It was found that migrant learners suffer from mental health problems ranging from anxiety, hostility, and fatigue compared to non-migrant learners.

One prospective study in human migration predicting migration destinations argued that the traditional migration models such as gravity and radiation, which are based on population and distance variables, are limited, and cannot deal with complicated migration dynamics [12]. The study used survey data and developed a simple decision tree and random forest model to predict the successful migration outcome. The simple decision tree provided better classification accuracy than the random forest. In a similar study conducted by [13], XGBoost (Extreme gradient boosting) and ANN (Artificial neural network) were used to model human migration. The study made similar observations where the XGBoost and ANN models outperformed traditional human models on a variety of assessment metrics. Robinson and Dilkina [13] argued that the traditional models have a fixed form and function and can only be used where a large amount of prior ground truth mobility data is unavailable.

The importance of quality education and the detrimental impact of learner migration on pedagogy and education administration cannot be overstated. Educational planners need to have the upper hand in understanding learner migration/mobility patterns and the underlying causes to inform planning and monitoring; and device sound policies to reduce uncertainties and negative impacts of learner migration. Although various qualitative and machine learning research has been conducted to understand factors contributing to learner migration, there is still a significant gap in the use of data mining techniques to assess learner migration and find optimum feature subsets more suited to modelling the phenomena.

The focus of this study is to find the most optimal feature subset to predict learner migration cases and related causes. This paper provides in-depth insights into the factors influencing learner migration, to improve educational planning, and promote the use of data analysis and machine learning to support data-driven decisions in the education system. The outline of the paper includes a literature review, theoretical framework, methods and materials, experiment design and execution. Furthermore, it presents the findings and discusses their implications and future work.

II. THEORETICAL FRAMEWORK

A. RAVENSTEIN MIGRATION THEORY

The information systems (IS) domain is the aspect of computing that focuses on the social context of technology [14]. Historically, most information systems research and data analysis have been conducted using an interpretive approach. The interpretive approach is not motivated by specific perspectives and has no logical boundaries, and the arguments are mainly based on the researchers' opinions and interpretations [14]. To avoid this, the study was anchored on the Ravenstein migration theory to guide the research and improve epistemological bases for confirming the validity and generalization of the results. In 1889, Ernest Ravenstein, widely known as one of the earliest theorists of migration, used census data from England and Wales to develop the laws of migration, which later became widely accepted among scholars [15]. The Ravenstein migration theory is determined by push-pull factors. The theory refers to push factors as unfavorable conditions from the point of origin, while pull factors are favorable conditions at the destination. The Ravenstein migration laws can be reformulated as outlined in Table 1:

Table 1. Ravenstein laws of migration [15]

The Ravenstein laws of migration
○ Most migrations are over short distances.
○ Migration happens in stages and as distance increases, the volume of migration decreases.
○ In general, long-range migrants move into urban areas.
○ Each migration produces a reverse movement, although not necessarily in the same volume.
○ Rural inhabitants are more migratory than urban inhabitants,
○ Women are more migratory than men within their own country, but males are more migratory over long distances.
○ Migrants are mostly adults. Families seldom migrate from their country of birth.
○ Large towns grow more through migration than through natural growth.
○ Migration increases with economic development.
○ Migration is mostly due to economic causes

Ravenstein's work has been widely accepted among academics and serves as the basis for many theories. One of these theories is that of Everett Lee, who extended Ravenstein's theory to give more emphasis to the push-pull factors. Lee [16] argues that a migration decision is influenced by factors related to place of origin, destination, intervening obstacles, as well as personal factors. Fig. 1 shows the Lee's migration model.

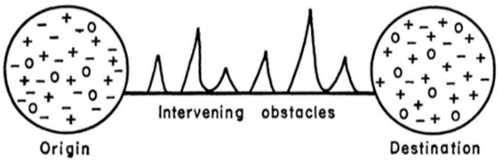


Figure 1. Lee's model of migration [16]

Lee further asserts that the migration process is selective for differentials such as gender, age, education, and social class; and they influence how one responds to push-pull factors, and the ability to overcome intervening obstacles. The research undertaken by [4, 5] in relation to the factors influencing learner migration aligns with Ravenstein's migration theory. These findings offer empirical evidence that supports the ongoing significance of Ravenstein's migration theory in understanding learner migration dynamics.

B. FEATURE SELECTION

Feature selection is a dimension-reduction technique that aims to reduce the input variable into a machine-learning model by removing irrelevant variables unrelated to the response variable [17–19]. Feature selection reduces the number of parameters in the model, the training time, and overfitting by improving generalization, and helps to avoid the curse of dimensionality [17]. Irrelevant variables consume processing capacities such as memory, time, cost, and other computational resources which negatively contribute to the outcome of the machine-learning exercise of some algorithms [17]. One needs to consider the fact that a feature that might be useful in one machine learning algorithm may be underrepresented or unused by another. In addition, it is still possible that a variable which shows little evidence of contributing to the explanation of the response variable may prove significantly useful in the presence of other variables [20]. The study used four widely accepted feature selection algorithms: J48, Boruta, Adaboost.M1, and RPART. These algorithms were chosen for their effectiveness in identifying informative features in various studies. The choice of algorithms used reflects a broad strategy in handling different aspects of the learner migration phenomenon. Their comparative analysis can lead to a comprehensive understanding of factors that influence learner migration. Boruta, as proposed by [21] is effective in identifying relevant features in a dataset with potentially many attributes.

Boruta has been successfully deployed in similar educational settings to predict student academic performance [22]. The study used features from students' demographic information, academic records, technological resources, social attitudes, family background, and socio-economic status which were subjected to feature selection. Boruta's performance was comparatively analyzed alongside Information Gain, ReliefF, and Recursive Feature Elimination. Performance metrics such as accuracy, kappa statistic, and f-measure were employed as a benchmark for the analysis. The findings demonstrated the effectiveness of the Boruta algorithm in reducing feature dimensionality, as it consistently outperformed other algorithms. RPART and J48 as discussed by [23, 24] are capable of providing insights into decision rules and patterns. Adaboost.M1 is effective, particularly in imbalanced datasets [25].

The present body of knowledge is not strong enough to trace the use of RPART, Adaboost.M1 and J48 in educational settings. One intriguing study is that of [26] in performing a comparative analysis of J48 and Adaboost.M1 (using J48 as a base classifier) FS techniques using WEKA (Waikato Environment for Knowledge Analysis) and open WEKA datasets (supermarket.arff, labor.arff, soybean.arff, and segment.arff). The study used several class labels in the dataset, accuracy, amount and length of the generated rules, error rate and standard deviation as the basis to assess performance. Multiple experiments were conducted, and the findings indicated that AdaBoost.M1 outperforms the J48 algorithm in terms of accuracy when the dataset contains exactly two class labels. On the other hand, the J48 demonstrates faster rule generation compared to AdaBoost.M1. However, when the dataset contains more than two class labels, the J48 algorithm performs better than AdaBoost.M1. There is a paucity of research where RPART is used for feature selection. A notable instance of where RPART was used is found in the work of [27] where the study comparatively analyzed the performance of C5.0, random forest, RPART, KNN, SVM and Boruta algorithms for feature selection in a cervical cancer prediction. [27] used Accuracy and AUC (Area Under Curve) metrics to assess the performance of the algorithms. The study favoured C5.0 and random forest classifiers as reasonably performing in identifying women exhibiting clinical signs of cervical cancer compared to other algorithms including RPART and Boruta.

This presents a novel opportunity for us to evaluate the performance of both Boruta and RPART in the educational setting. In addition, to assess multicollinearity and validate the importance scores obtained from the four algorithms, we used mutual information and Spearman correlation. These techniques provided insight into the relationships between the selected features and helped confirm the robustness of the feature importance rankings. To evaluate the performance of the four FS algorithms, we used Cohen's kappa, accuracy, and standard deviation. These metrics allowed us to assess the reliability and stability of the feature selection process [22, 27]

C. BORUTA

Boruta is a feature selection algorithm that acts as a wrapper for a random forest classifier [28]. Boruta derives its name from a demon in Slavic mythology that lived in pine forests [29]. The random forest as an ensemble in Boruta is often used in classification or regression problems and can handle high-dimensionality feature selection problems [17, 30, 31]. The random forest builds decision trees on different samples of the data and uses the majority vote for classification and average in case of regression problems [28, 30]. During the feature selection process, Boruta adds randomness to the given data set by creating mixed copies of all features (referred to as shadow features). It then trains a random forest classifier on the extended data set and applies a feature importance measure (the default is Mean Decrease Accuracy) to rank the importance of each feature. The higher mean decrease in accuracy suggests the relevance of the feature. At each iteration, it checks whether a true feature has higher importance than the best of its shadow features (i.e., whether the feature has a z-score higher than the maximum z-score of its shadow features) and continuously removes features that are deemed unimportant. The algorithm will stop when all features are confirmed or rejected, or it reaches a certain limit of random forest runs [21].

D. RPART

Recursive Partitioning (RPART) is an open-source implementation of CART [24]. The RPART is a decision tree that constructs classification or regression models on a general structure using a two-phased process. In phase one, the single variable that best separates the data into two groups is found. The data is separated, and this process is applied recursively to each subgroup separately until a predetermined termination criterion is met [24, 32]. At each step, the split is based on the predictor variable that resulted in the largest possible reduction in heterogeneity of the response variable. Splitting rules can be built in many various ways, all of which are based on the concept of impurity which is a measure of the degree of heterogeneity of the tree leaf nodes. In phase two, the algorithm uses cross-validation to trim back the full tree, compute risks in all the sub-trees constructed, and finally choose the one with the lowest estimate of risk. The RPART uses the GINI Index and Entropy as possible parameters to quantify the impurity of the leaf nodes [33].

E. ADABOOST

AdaBoost (Adaptive Boosting) is an ensemble or meta-algorithm used to improve the classification performance of weak classifiers [25]. It first creates a set of poor learners/baseline classifiers where all features in the training data are weighted equally. Then the weight of incorrectly classified features is increased while the weight of correctly classified features is decreased. This process runs until complete classifier sets are created. It will then initiate a voting process, with each weak classifier giving a weighted vote to make a classification decision [25, 34].

F. J48

The J48 algorithm is the most used and works accurately for many classification problems, both for categorical and continuous data [23]. J48 is a decision tree classifier that works hierarchically, with each level representing a feature. It uses a C4.5 that uses information gain to select features at each stage [23, 25]. The algorithm first selects the function to partition the training data into a class using information gain. The tree either immediately classifies the data or moves it to the next level of the tree depending on the value of that feature. The process is repeated iteratively until all training data is classified. The performance of the final tree is then evaluated against test data [23, 25].

G. SPEARMAN CORRELATION

Spearman's correlation is a non-parametric bivariate test that measures the strength of the association between two variables [35]. In contrast to Pearson, who measures a linear relationship, Spearman measures a monotonic relationship between two variables [36, 37]. The relationship is said to be monotonic when the variables move together in the same direction, but not necessarily at a constant rate as in linear relationships [36]. The Spearman correlation gives a correlation coefficient that varies between “+1” and “-1”. Correlation coefficients of “+1” or “-1” indicate a perfect degree of association between the two variables [36]. When the value of the correlation coefficient approaches 0, the relationship between the two variables is said to be weaker [35, 36]. The direction of the relationship is indicated by the sign of the coefficient, with the “+” sign indicating a positive relationship and the “-” sign indicating a negative relationship. Spearman can be calculated mathematically as follows [36]:

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}, \tag{1}$$

where n – number of pairs and d – difference in number of pairs

Spearman generates a p-value (probability value) that indicates how likely it is that the results of the relationship (correlation coefficient) arose by chance [38]. The p-value is mainly used to either accept or reject the null hypothesis; The smaller the p-value, the stronger the evidence to reject the null hypothesis (there is no relationship) and to accept the hypothesis (there is a relationship). Statistically, the p-value < 0.05 is accepted to reject the null hypothesis [38]. Although there is no consensus in the literature [35, 36] to assess the strength of the relationship based on the correlation coefficient, the following division is generally accepted. A general guide to interpreting the strength of a relationship based on the correlation coefficient is presented in Table 2 [35]

Table 2. Correlation strength scale

Absolute Value of Coefficient	Strength of Relationship
r < 0.3	None or very weak
0.3 < r <0.5	Weak
0.5 < r < 0.7	Moderate
r > 0.7	Strong

H. MUTUAL INFORMATION

Mutual information (MI) provides an alternative to calculate the association between two variables [39, 40]. In contrast to Spearman correlation analysis, which provides a quantitative means of measuring the strength and direction of the association, mutual information calculates the amount of knowledge about one variable that can be obtained simply from knowing the value of another variable [41]. Mutual information takes a value between “0” and “1”. A large reduction in uncertainty is indicated by high mutual information and vice-versa. When the variables are independent their value is likely to be zero [39, 40]. It is safe to point out that when variables share any large amount of data, the mutual information can sometimes have an upper bound greater than 1. This is high levels of disorder and entropy (low levels of purity) [41] which increases the uncertainty of a random variable [41, 42]. The formula for calculating mutual information is illustrated below:

$$I(x{:}\,y) = \sum_{i=0}^{n} \square \sum_{j=1}^{n} p(x(i), y(j)).\log\left(\frac{p(x(i), y(j))}{p(x(i)).p(y(j))}\right), \quad (2)$$

where *MI*=0 when *x* and *y* are statistically independent. The MI is related linearly to entropies of the variables through the following formula.

$$I(x{:}\,y) = \begin{cases} H(x) - H(x|y) & \textit{if } x \textit{ and } y \textit{ are independent} \\ H(y) - H(y|x) & \textit{if } y \textit{ and } x \textit{ are independent} \\ H(x) + H(y) - H(x,y) & \textit{otherwise} \end{cases}, \quad (3)$$

where *(x; y)* is the mutual information for *x* and *y*; *H(x)* is the entropy for *x; H(y)* is the entropy for *y; H (x|y)* is the conditional entropy for *x* given *y* and *H (y|x)* is the conditional entropy for *y* given *x*.

Although there is no consensus in the literature to assess the strength of the relationship based on the information value, the following split (Table 3) is more generally accepted.

Table 3. Mutual information strength scale

Information Value	Predictive power
<0.02	Useless
0.02 to 0.1	Weak predictors
0.1 to 0.3	Medium Predictors
0.3 to 0.5	Strong predictors
>0.5	Suspicious

III. METHODS AND MATERIALS

In this study, a conventional data mining technique, commonly referred to as Knowledge Discovery in Databases (KDD), was used. The process includes several important steps. Initially, data was assembled and preprocessed/cleaned to ensure the quality and relevance of the dataset. We then conducted exploratory data analysis to gain a preliminary understanding of the underlying structure of the data. The data was then transformed into relevant variables ready for the machine learning exercise. Machine learning feature selection algorithms were used to create predictive models and feature importance scores. The models were then evaluated using an assortment of model performance metrics to ensure their accuracy and reliability. Finally, the results were discussed and used to draw meaningful conclusions that contributed to a deeper understanding of the research objective [43].

A. DATA ASSEMBLY

The EMIS database consists of over 400 tables and thousands of attributes describing learners, educators, and school inventory. The Ravenstein migration theory was used to guide the initial data assembly for the study. Features that broadly conform to the principles in the Ravenstein migration theory were selected and are outlined in Table 4.

Table 4. Features selected for study.

Code	Feature	Description
f01	learner_age_at_entry	Learner age at the time of first admission
f02	learner_age_at_exit	Learner age when leaving the school
f03	school_boarding_facilities	Identifies if a learner is in a school hostel or not
f04	learner_deceased_parent	Identifies if a learner has deceased parents
f05	learner_gender	Learner Gender
f06	learner_current_grade	Grade of a learner
f07	learner_years_in_school	Number of years a learner is admitted to the school
f08	school_exit_grade	The last grade of the schools
f09	learner_grade_years	Number of years a learner was in a grade
f10	learner_home_language	The home language of a learner
f11	school_instruction_language	School language of learning and teaching
f12	learner_lsen_status	Identifies if a learner has a disability or not
f13	learner_phase_years	Number of years in a phase
f14	learner_preferred_language	Language a learner prefers to be taught with
f15	learner_progressed	Identifies if a learner was progressed/condoned to a grade or not
f16	school_psnp	Identifies if a learner is benefitting from the school nutrition programme
f17	learner_race	Race of a learner
f18	learner_transport	Type of the transport a learner is using to school
f19	school_district	District of the school a learner is enrolling
f20	school_type	Type of the school a learner is enrolling
f21	school_sector	A sector of the school a learner is enrolling
f22	school_quintile	Poverty ranking of the school a learner is enrolling
f23	school_promotion_rate	Average school performance
f24	learner_displacement_count	Frequency of displacement
c25	movement_indicator	response class to assess the possibility of learner displacement (1= “displacement”. 2= “no displacement”)

B. DATA PREPARATION

Equally important, data preparation is another important phase of the machine learning exercise. It is always difficult to get the data ready for mining. Therefore, a lot of time has been devoted to this exercise, being careful not to tamper with the underlying knowledge and structure of the data. Most categorical variables contained non-standard attributes, some of which have been fixed and others removed. Records of learners that we could not positively identify when transitioning through the system were removed. Several attributes for example, *f01, f02, f07, and f24* were not part of the original data but were derived from other features contained in the data set. Feature like *f23* resulted from merging with other learner performance-related features to reduce attribute complexity. Duplicate records in the data were also identified and removed.

C. SAMPLING METHOD

A sample of 10% (12849 observations and 25 features) was extracted from the cleansed data using a simple random sampling technique. This sample represents the operational school data between 2011 and 2016 in atomic format. The sampling method used is considered a basic method of sampling, where each member of a population has the same chance of being included in the sample. Machine learning algorithms are sensitive to data with an imbalance distribution of response class as they are forced to create classifiers that are biased toward the majority class and lead to increased generalization error [44]. Having said that, the response class was tested for sensitivity to assess if there were any problems in the distribution of the class and the performance of the sample. The data was partitioned into train and test data using 70% train and 30% test split respectively. The random Forest classifier was used to predict the response class in the test data and the following performance information was generated.

Table 5. Sample performance metrics

Confusion matrix	Other performance metrics
## Reference	##Sensitivity: 1.000
## Prediction 1 2	##Specificity: 1.000
## 1 1896 0	##Pos Pred Value: 1.000
## 2 0 1920	##Neg Pred Value: 1.000
##Accuracy: 1	##Prevalence: 0.497
##95% CI: (0.999, 1)	##Detection Rate: 0.497
##No Information Rate: 0.5031	##Detection Prevalence: 0.497
##P-Value [Acc > NIR]: < 2.2e-16	##Balanced Accuracy: 1.000
## Kappa: 1	

The model correctly recognized all observations. The accuracy, kappa, and sensitivity of the model were optimal, which suggested that the model can recognize all classes equally [45] This further suggested that the sample would perform well in the experiment. The findings are interesting and imply that there is a promise in using a random forest to address the objectives of the study.

D. EXPERIMENT DESIGN

The sampled data was processed using the caret package in RStudio. The data was subjected to Boruta (*a wrapper for random forest),* RPART, J48, and Adaboost machine learning algorithms using cross-validation techniques (*10-fold cross-validation)* to assess features that have the potential to predict a response class (*c25≈ movement indicator*). We then subjected the same sample data to Spearman's correlation, and mutual information to generate a relationship matrix and associated p-values for further analysis.

IV. RESULTS

A. OPTIMUM VARIABLES

This section illustrates the results of the feature selection (FS) methods. The performance of features varied across different algorithms (Fig. 2 & Table 6). In general, there is a significant degree of agreement among the algorithms in ranking the features.

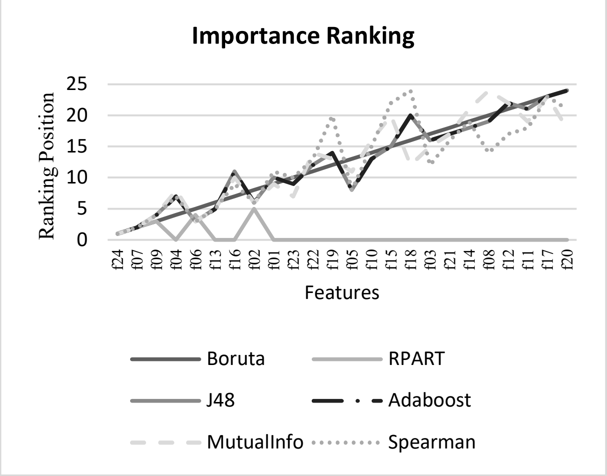


Figure 2. Importance Ranking

While there are partial inconsistencies in the subsets of variables used between different feature selection methods, features such as f06, f07, f09, and f24 have relatively high rankings and appear in the top 5 across all feature selection algorithms (Table 6).

Table 6. Variable importance scores and ranking

	Machine Learning				Statistical			Calculated Ranking					
Code	Boruta	RPART	J48	Adaboost	MutualInfo	Spearman	Spearman (p-values)	Boruta	RPART	J48	Adaboost	MutualInfo	Spearman
f24	102.68	100.00	100.00	100.00	0.69	0.97	0.00	1	1	1	1	1	1
f07	27.76	41.12	56.20	56.20	0.27	-0.50	0.00	2	2	2	2	2	2
f09	21.90	19.59	43.47	43.47	0.10	0.43	0.00	3	3	4	4	4	4
f04	19.70	0.00	23.33	23.33	0.03	0.23	0.00	4	0	7	7	8	7
f06	14.63	15.80	49.57	49.57	0.11	-0.46	0.00	5	4	3	3	3	3
f13	14.13	0.00	37.65	37.65	0.07	-0.34	0.00	6	0	5	5	5	5
f16	14.03	0.00	6.27	6.27	0.01	-0.10	0.00	7	0	11	11	10	9
f02	12.96	10.86	33.73	33.73	0.07	-0.30	0.00	8	5	6	6	6	6
f01	12.47	0.00	9.31	9.31	0.01	0.08	0.00	9	0	10	10	9	11
f23	11.05	0.00	10.34	10.34	0.03	0.10	0.00	10	0	9	9	7	10
f22	7.48	0.00	4.52	4.52	0.00	0.04	0.00	11	0	12	12	14	13
f19	7.35	0.00	1.46	1.46	0.00	0.01	0.14	12	0	14	14	13	20
f05	6.89	0.00	12.69	12.69	0.01	0.13	0.00	13	0	8	8	11	8
f10	6.40	0.00	2.50	2.50	0.00	0.02	0.01	14	0	13	13	16	15
f15	5.83	0.00	1.02	1.02	0.00	-0.01	0.19	15	0	15	15	20	22
f18	5.62	0.00	0.33	0.33	0.00	0.00	0.71	16	0	20	20	12	24
f03	5.55	0.00	1.00	1.00	0.00	0.05	0.00	17	0	16	16	15	12
f21	4.74	0.00	0.67	0.67	0.00	0.02	0.02	18	0	17	17	17	16
f14	4.36	0.00	0.42	0.42	0.00	-0.01	0.14	19	0	18	18	21	19
f08	2.75	0.00	0.33	0.33	0.00	-0.04	0.00	20	0	19	19	24	14
f12	2.62	0.00	0.12	0.12	0.00	-0.02	0.08	21	0	22	22	22	17
f11	1.97	0.00	0.13	0.13	0.00	-0.02	0.08	22	0	21	21	19	18
f17	1.39	0.00	0.01	0.01	0.00	0.00	0.68	23	0	23	23	23	23
f20	0.69	0.00	0.00	0.00	0.00	-0.01	0.15	24	0	24	24	18	21
std	19.85	21.39	24.60	24.60	Standard Deviation								

This suggests that these features are strongly correlated or associated with the target variable (*c25*). Similarly, f02, f04, f013, and f23 consistently appear in between the 5th and 10th position across all algorithms except for RPART which only selected five features. Learner Gender (f05) has notably different rankings, with J48, Adaboost, and Spearman placing it in the 8th position, while MutualInfo and Boruta ranked it relatively lower at the 11th and 13th position respectively. Learner age at entry (f01), f16 (*school nutrition project or PSNP*) and f22 (*school quintile*) have relatively inconsistent rankings across the algorithms, showing variations between the 7th -14th position. Interestingly, the results and variable importance scores for Adaboost and J48 are identical (Table 6). Boruta rejected two features, f17 (*learner race*) and f20 (*school type*). The Learner race (f20) was also confirmed unimportant by Adaboost and J48.

B. FILTER-BASED METHOD (STATISTICAL METHODS)

The relationship of the 24 predictor variables with the response variable was further explored using statistical methods. Spearman's correlation coefficients ranged between "*0*" to "*0.97*" indicating diversity, variability, and consistency in the identification of the priority variables. Table 6 illustrates the importance scores and derived rankings of the features.

It is interesting that features (*f06, f07, f09, f24*) which appear in the top 5 of J48, Adaboost, Boruta, and RPART also appear in the top 5 of both spearmen and mutual information, with a correlation coefficient above *"0.3"* and information value above *"0.02"* respectively (Table 6). Fig. 3 illustrates the variable importance scores of the four machine learning algorithms against the Spearman and mutual information.

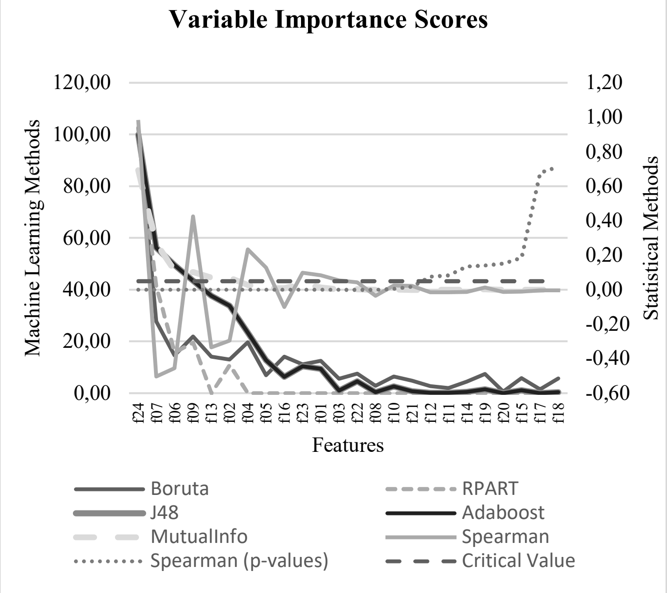
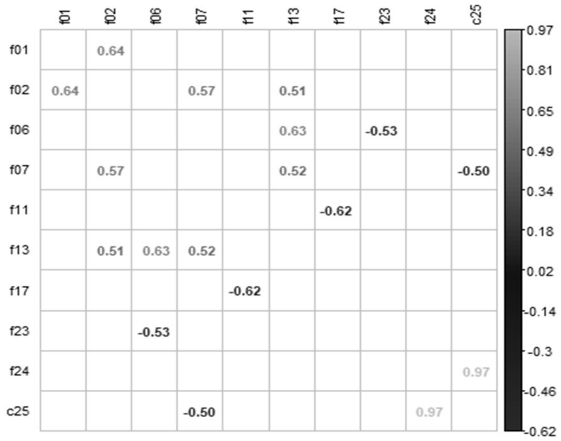


Figure 3. Feature importance scores of Boruta, J48, RPART, and Adaboost.M1 against the spearman and mutual information

The p-values for the four features are "0", giving us some confidence in the stability of their predictive ability. Another observation is that when there are large disparities among the variable importance of the machine learning FS methods, the Spearman p-values increase. Similarly, most features that were ranked lower by the machine-learning FS methods were also ranked lower by the statistical methods, and the Spearman p-values are also above the acceptable critical value threshold of 0.05 (Table 6). The Spearman p-value increases above the acceptable critical value when the importance score of J48 and Adaboost drops below 1.5 except for f03 and f08.

C. CLUSTERS, CORRELATIONS & MULTICOLLINEARITY

The matrix in Fig. 4 shows all Spearman correlation coefficients above 0.49. From the correlation matrix, we can rule out the problem of multicollinearity. Multicollinearity happens when two or more predictor variables are linearly related.



	f01	f02	f06	f07	f11	f13	f17	f23	f24	c25
f01		0.64								
f02	0.64			0.57		0.51				
f06						0.63		-0.53		
f07		0.57				0.52				-0.50
f11							-0.62			
f13		0.51	0.63	0.52						
f17					-0.62					
f23			-0.53							
f24										0.97
c25				-0.50					0.97	

Figure 4. Spearman Correlation Matrix (correlation coefficient >0.49 ≈ very weak to a strong relationship. Refer to Table 2 for strength interpretation)

In general, an absolute correlation coefficient > 0.7 between two or more predictors signifies the existence of multicollinearity. Multicollinearity is always a concern in classification or regression problems since it has the potential to undermine the statistical significance of the response variables.

Furthermore, we have evaluated whether there exist any associations among data points of distinct. The associations can either be in the form of clusters or correlations. This can assist in gaining insights into the underlying relationships or similarities between the features. Clustering occurs when a collection of data points closely aligns while being relatively distant from data points residing in other clusters. When the data points form a cluster, it means that those data points share similar values or characteristics, and therefore suggest that there might be a correlation, association, or similarity between the values of those features. Data points are correlated when there exists a linear relationship or dependence between their values. Fig. 5 shows possible clustering, correlations, or associations of the distinct features.

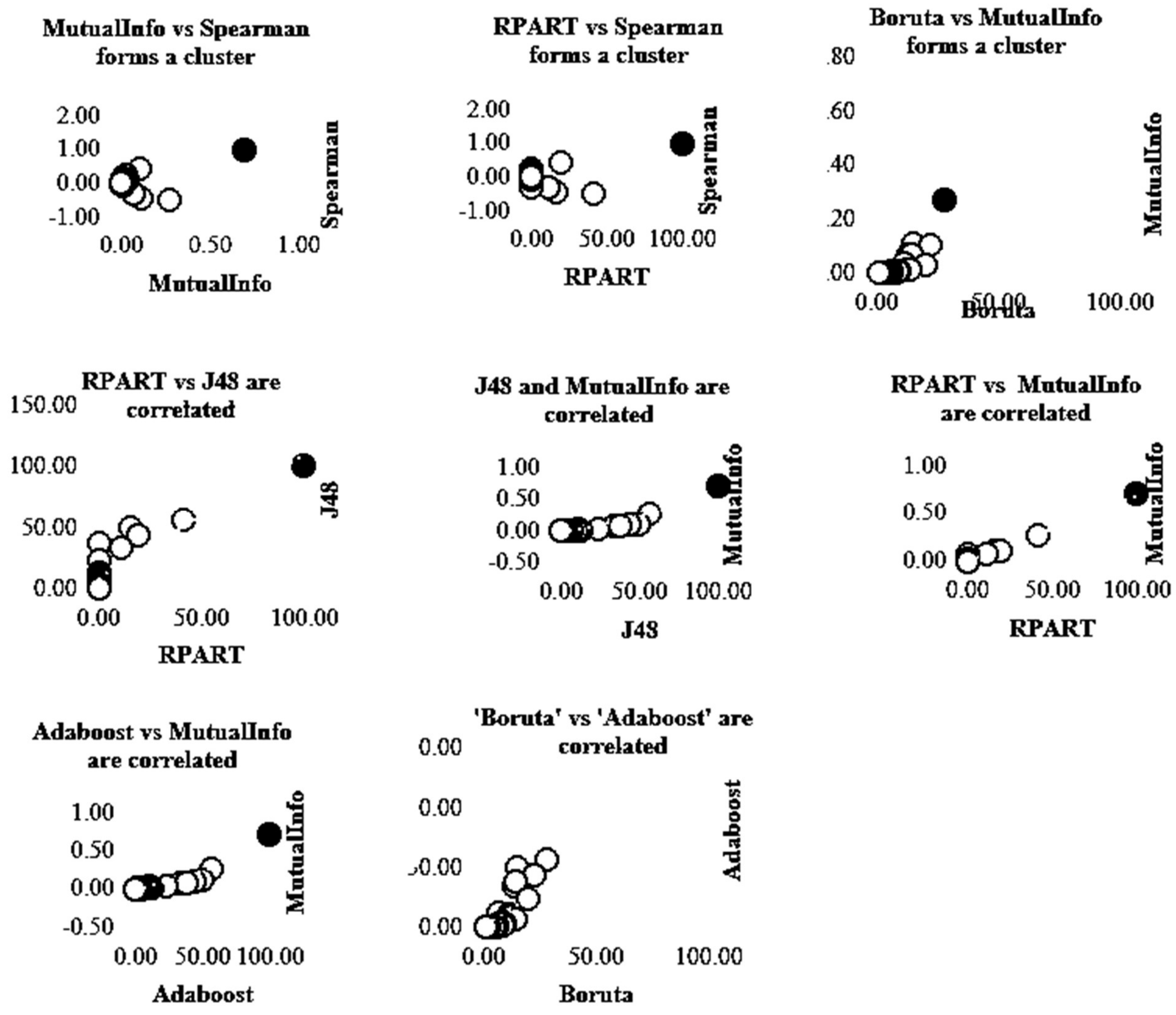


Figure 5. Variable importance scores and ranking. We have compared the rankings of the algorithms as depicted in Fig. 2 and the correlations presented in Table 6, and this provides empirical evidence that the algorithms assume a shared understanding of the data structure.

D. Model Evaluation Metrics

The feature selection methods were configured to use 10-fold cross-validation repeated 10 times. There are several points to note concerning the results. According to the model performance metrics, both Adaboost, J48, and RPART performed well (Table 7). The results show that the RPART model started with a simpler tree with cp=1.0, accuracy, and kappa of “0.5068” and “0” respectively. The RPART added complexity to the model by reducing the cp and resolved the model with Mallow's cp=0.5 (complexity parameter), accuracy =1, and Cohen’s Kappa=1 which shows high classification accuracy. Similarly, the performance of the J48 model was resolved with c=0.01(pruning confidence), m=1(minimum instances), accuracy =1, and Cohen’s Kappa=1. It is interesting that the Adaboost was switched between two base classifiers (Adaboost.M1 and Real Adaboost) while trying to increase the number of iterations (nIter). However, the performance of the model did not change. The model’s performance was resolved using Adaboost.M1 with nIter=50, accuracy =1, and Cohen’s kappa=1.

Table 7. Model Performance Metrics

```
## C4.5-like Trees
## 12849 samples, 24 predictors
## 2 classes: '1', '2'
## No pre-processing
## Resampling: Cross-Validated (10-fold, repeated 10 times)
## Summary of sample sizes: 11564, 11564, 11564, 11564, 11564,
1565, ...
## Resampling results across tuning parameters:
```

C	M	Accuracy	Kappa
0.010	1	1	1
0.010	2	1	1
0.010	3	1	1
0.255	1	1	1
0.255	2	1	1
0.255	3	1	1
0.500	1	1	1
0.500	2	1	1
0.500	3	1	1

*Accuracy was used to select the optimal model using the largest value. The final values used for the model were C = 0.01 and M = 1.

```
## CART (RPART)
## 12849 samples, 24 predictors
## 2 classes: '1', '2'
## No pre-processing
## Resampling: Cross-Validated (10-fold, repeated 10 times)
## Summary of sample sizes: 11564, 11564, 11564, 11564, 11564,
11565, ...
## Resampling results across tuning parameters:
  cp      Accuracy   Kappa
  0.0     1.0000     1
  0.5     1.0000     1
  1.0     0.5076     0

Accuracy was used to select the optimal model using the largest
value. The final value used for the model was cp = 0.5.
```

```
#AdaBoost Classification Trees
## 12849 samples, 24 predictors
## 2 classes: '1', '2'
## No pre-processing
## Resampling: Cross-Validated (10-fold, repeated 10 times)
## Summary of sample sizes: 11564, 11564, 11564, 11564, 11564,
11565, ...
## Resampling results across tuning parameters:
  nIter   Method                  Accuracy       Kappa
  50      Adaboost.M1                  1            1
  50      Real Adaboost                1            1
  100     Adaboost.M1                  1            1
  100     Real Adaboost                1            1
  150     Adaboost.M1                  1            1
  150     Real Adaboost                1            1

*Accuracy was used to select the optimal model using the largest
value. The final values used for the model were nIter = 50 and
method = Adaboost.M1
```

V. DISCUSSION

The feature selection methods presented in this work are general and applied in many feature selection exercise. In this section, we present an overview of the literature related to the presented work to support the selection of the optimal feature subset. The three FS methods returned *accuracy =1* and *Cohen's kappa =1* respectively. The literature [46, 47] describes the accuracy metric as a measure of how generally the model has performed across all the classes. It provides a percentage of correctly classified instances from all instances [47]. On the other hand, Cohen's kappa measures agreement between interrater on the ground truth labels versus model predictions. It accounts for the imbalance in the class distribution as opposed to calculating the overall accuracy percentage. It represents the extent to which the data used correctly represent the measured variables. Both accuracy and Cohen's kappa range from *0 to 1, with 1 = 100%* indicating excellent model performance [47]. The accuracy metric has limitations especially when working on data with an imbalance class distribution [46]. It was therefore not considered a reliable measure of performance for the models being studied. The literature [48] argues that kappa is more useful than accuracy especially when dealing with class imbalance data. Cohen kappa can be interpreted as follows (Table 8):

Table 8: Kappa Interpretation

kappa	Predictive power
≤ 0	no agreement
0.01–0.20	none to slight
0.21–0.40	fair
0.41– 0.60	Moderate
0.61–0.80	substantial
0.81–1.00	almost perfect agreement.

From the results, we can therefore deduce that three FS methods (J48, RPART, and Adaboost.M1) have so far performed well with sufficient accuracy and reliability. We will then proceed and interpret what parameters say about model performance.

The three models (Adaboost, J48, and RPART) run with different parameters to maximize classification accuracy. An interesting parameter is that of RPART, called the complexity parameter (*cp*). RPART uses *cp* to avoid overfitting the model and save computation time [49, 50]. The *cp* is used to trigger the model's stopping rules such that the relative error reduction resulting from the best split falls below the *cp*. The larger *cp* values result in a higher penalty and produce a smaller tree with missing predictor variables. When the latter happens, the model finds a cross-validated error with the *cp* value, which can offer an optimal tradeoff between minimizing the misclassification error and the complexity of the tree depth. RPART implements a minimal optimal approach. On the other hand, the J48 uses pruning confidence (*c*) and minimum instances (*m*) parameters to maximize model accuracy and create simpler trees [51]. The confidence factor/pruning confidence provides a threshold of allowable inherent error in the data while pruning the decision tree. The lower threshold increases pruning and consequently generates more general models (*models with high classification performance*). To get simpler models, the minimum instances can be adjusted, where a lower number means a simpler model [51]. This better explains why J48 chose c = 0.01 (pruning confidence) and m = 1 (minimum instances) to resolve the model accuracy (Table 7).

Boruta works differently from J48, Adaboost, and RPART. It implements a novel feature selection algorithm to find all relevant variables proved important by the statistical tests [21, 28]. In Boruta, features do not compete among themselves like in J48, Adaboost, and RPART, but with a randomized version of themselves [52]. Though the classification accuracy of J48, Adaboost, and RPART are high, the literature [52] asserts that one cannot rely on classification accuracy or Cohen's kappa as a criterion for selecting features as important or rejecting as unimportant. It argues that the reduction of classification accuracy upon removal of the feature is sufficient to declare the variable important, but the absence of this effect is not enough to consider the feature unimportant [52].

Spearman correlation and mutual information are non-parametric (*distribution-free*) tests that provide a simple approach to feature selection. These methods enabled the process of ruling out multicollinearity among predictor variables, which has the potential to undermine the statistical significance of the response variables. In addition, the statistical methods assisted in confirming the results of machine learning algorithms. However, they have their share of limitations. Since they are assessing the relationship of two variables at a time, they lack enough proof that the feature cannot be important in conjunction with other features [52]. In addition, Spearman correlation and mutual information can detect linear or monotonic relationships accurately but fail to detect quadratic relationships if present in the variable space.

It is common practice in feature selection to subject all available features to various FS algorithms before deciding on the optimal feature subset. However, it is safe to point out that different FS learn the relationship between the predictor and the response variables differently. In principle, the variables most used by various FS algorithms are considered the most important [20]. Given the latter, it is tempting to choose the RPART feature subset as optimal since it meets all accuracy requirements discussed and supported by the literature and its minimalist approach. What will be the cost of such a decision?

VI. FUTURE WORK

The performance of the sample was tested using a random forest classifier. It's worth noting that Boruta also functions as a wrapper around the random forest classifier [28]. This inherent relationship could potentially impact the ultimate findings of the study. Further study is needed to assess the influence of this relationship. The results of Adaboost.M1 and J48 provided a very interesting outcome due to the similarity in importance scores for all features. One possible explanation for this incident is that the Adaboost.M1 algorithm utilized in R Studio is an upgraded variant that employs the j48 algorithm as its underlying classifier [53]. The research by Eibl and Pfeiffer [54] discovered that Adaboost.M1 does not work with too weak classifiers and recommended a few changes to the algorithm to produce a variation called Adaboost.M1W. More research is needed to understand these results.

VII. CONCLUSION

In this paper, we mainly compared four types of machine learning FS methods (J48, Adaboost.M1, RPART, and Boruta) in finding optimal feature set to predict learner movement/displacement possibilities in the Limpopo education system. In addition, the two statistical tests (Mutual Information and Spearman Correlation) were used to validate the results of the FS algorithms and rule out multicollinearity among predictor variables. The results demonstrated that all the machine learning FS methods used different variables, and all performed well in terms of accuracy, Cohen's kappa, specificity, and sensitivity. RPART used a minimal optimal approach, and this makes it a more attractive choice. However, it excluded most of the variables confirmed by literature [4–6] and the theoretical framework [14–16] as core drivers for the migration phenomenon. The phenomenon of human migration is inherently complex, and one would need to understand all the contributing variables related to learner migration. J48, Adaboost.M1, and Boruta provide alternative choices due to their performance. Table 6 shows the standard deviations of feature sets per algorithm.

The literature asserts that lower standard deviation or variance is an important factor in feature selection methods. When the standard deviation of a feature subset is small, it indicates that the selected features have relatively low variability or spread, which suggests that the chosen features will be more consistent and stable across different samples. This can lead to a simpler and more interpretable model, as it focuses on the most stable and reliable features, which can mitigate the risk of overfitting and enhance the model's performance [55]. Given the latter, Boruta with 22 confirmed features, and a relatively lower standard deviation of 19.85 becomes the ultimate feature selection for this study. In addition, the feature sets in Boruta are in line with traits that have been supported by the literature [4–6] and the Ravenstein theory of migration [14–16] which underpins this study.

REFERENCES

[1] A. Sîrbu *et al.*, "Human migration: The big data perspective," *Int J Data Sci Anal*, vol. 11, no. 4, pp. 341–360, 2021, https://doi.org/10.1007/s41060-020-00213-5.

[2] R. J. Lennox *et al.*, "Conservation physiology of animal migration," *Conserv Physiol*, vol. 4, no. 1, pp. 1–15, 2016, https://doi.org/10.1093/conphys/cov072.

[3] F. Chiororo, "Leadership for learning in Zimbabwean secondary schools: Narratives of school heads (Publication No. 18905)," Doctoral Thesis, University of Kwa-Zulu Natal, Durban, 2020. Accessed: Aug. 25, 2023. [Online]. Available at: https://researchspace.ukzn.ac.za/handle/10413/18905

[4] R. J. Botha and T. G. Neluvhola, "An investigation into factors that contribute to learner migration in South African schools," *The Journal of Social Sciences Research*, vol. 6, no. 63, pp. 224–235, 2020, https://doi.org/10.32861/jssr.63.224.235.

[5] I. C. Simelani, "Learner migration and its Impact on rural schools : A case study of two rural schools in Kwazulu- Natal (Publication No.12637)," Masters Thesis, University of Kwazulu-Natal, Durban, 2016. Accessed: Aug. 25, 2023. [Online]. Available at: https://researchspace.ukzn.ac.za/handle/10413/12637

[6] H. Van der Merwe, "Migration patterns in rural schools in South Africa: Moving away from poor quality education," *Education as Change*, vol. 15, no. 1, pp. 107–120, 2011, https://doi.org/10.1080/16823206.2011.576652.

[7] H. Hanna, "Being a migrant learner in a South African primary school: Recognition and racialisation," *Child Geogr*, pp. 1–16, 2022, https://doi.org/10.1080/14733285.2022.2084601.

[8] G. Tati, "Student migration in South Africa," *Espace Popul Soc*, vol. 2, no. 3, pp. 281–296, 2010, https://doi.org/10.4000/eps.4160.

[9] A. K. Hallberg, "Student migration aspirations and mobility in the global knowledge society: The case of Ghana," *Journal of international Mobility*, vol. 7, no. 1, pp. 23–43, 2020, https://doi.org/10.3917/jim.007.0023.

[10] A. Algarni, "Data mining in Education," *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 7, no. 6, pp. 58–77, 2016, https://doi.org/10.4018/978-1-5225-1877-8.ch005.

[11] L. Juan, "Analysis of the Mental Health of Urban Migrant Children Based on Cloud Computing and Data Mining Algorithm Models," *Sci Program*, vol. 2021, 2021, https://doi.org/10.1155/2021/7615227.

[12] S. M. R. Islam, N. N. Moon, M. M. Islam, R. A. Hossain, S. Sharmin, and A. Mostafiz, "Prediction of migration outcome using machine learning," *Springer link, International Conference on Deep Learning, Artificial Intelligence and Robotics*, vol. 441, pp. 169–182, 2022, https://doi.org/10.1007/978-3-030-98531-8_17.

[13] C. Robinson and B. Dilkina, "A machine learning approach to modeling human migration," *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, COMPASS 2018*, no. 1, 2018, https://doi.org/10.1145/3209811.3209868.

[14] S. Gregor, "A Theory of Theories in Information Systems," *Information Systems Foundations*, pp. 1–18, 2002. https://doi.org/10.3127/ajis.v10i1.439.

[15] D. B. Grigg, "E . G . Ravenstein and the ' laws of migration ,'" *J Hist Geogr*, vol. 3, no. 1, pp. 41–54, 1977. https://doi.org/10.1016/0305-7488(77)90143-8.

[16] E. S. Lee, "A Theory of Migration," *Demography*, vol. 3, no. 1, pp. 47–57, 1996, Accessed: Sep. 04, 2023. https://doi.org/10.2307/2060063.

[17] R. C. Chen, C. Dewi, S. W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *J Big Data*, vol. 7, no. 1, pp. 7–52, 2020, https://doi.org/10.1186/s40537-020-00327-4.

[18] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Adv Bioinformatics*, vol. 2015, no. 1, pp. 1–13, 2015, https://doi.org/10.1155/2015/198363.

[19] J. Miao and L. Niu, "A survey on feature selection," *Procedia Comput Sci*, no. 91, pp. 919–926, 2016, https://doi.org/10.1016/j.procs.2016.07.111.

[20] S. Velliangiri, S. Alagumuthukrishnan, and S. I. Thankumar Joseph, "A review of dimensionality reduction techniques for efficient computation," *Procedia Comput Sci*, vol. 165, pp. 104–111, 2019, https://doi.org/10.1016/j.procs.2020.01.079.

[21] M. B. Kursa, A. Jankowski, and W. R. Rudnicki, "Boruta – A system for feature selection," *Fundam Inform*, vol. 101, no. 4, pp. 271–285, 2010, https://doi.org/10.3233/FI-2010-288.

[22] P. Thereza, G. Lumacad, and R. Catrambone, "Predicting Student Performance Using Feature Selection Algorithms for Deep Learning Models," *Proceedings of the 2021 XVI Latin American Conference on Learning Technologies (LACLO)*, 2021, pp. 1–7. https://doi.org/10.1109/LACLO54177.2021.00009.

[23] N. Saravanan and V. Gayathri, "Performance and classification evaluation of J48 algorithm and kendall's based J48 algorithm (KNJ48)," *International Journal of Computer Trends and Technology*, vol. 59, no. 2, pp. 73–80, 2018, https://doi.org/10.14445/22312803/IJCTT-V59P112.

[24] C. Strobl, J. Malley, and G. Tutz, "An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests," *Psychol Methods*, vol. 14, no. 4, pp. 323–348, 2009, https://doi.org/10.1037/a0016973.

[25] J. Smucny, I. Davidson, and C. S. Carter, "Comparing machine and deep learning-based algorithms for prediction of clinical improvement in psychosis with functional magnetic resonance imaging," *Hum Brain Mapp*, vol. 42, no. 4, pp. 1197–1205, 2021, https://doi.org/10.1002/hbm.25286.

[26] P. Pandey and R. Prabhakar, "An analysis of machine learning techniques (J48 & AdaBoost)-for classification," *Proceedings of the 2016 1st India International Conference on Information Processing (IICIP)*, 2016, pp. 1–6. https://doi.org/10.1109/IICIP.2016.7975394.

[27] B. Nithya and V. Ilango, "Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction," *SN Appl Sci*, vol. 1, no. 6, 2019. https://doi.org/10.1007/s42452-019-0645-7.

[28] L. Breiman, "Random forests," *International Journal of Advanced Computer Science and Applications*, no. 6, pp. 1–33, 2001, https://doi.org/10.14569/IJACSA.2016.070603.

[29] F. Degenhardt, S. Seifert, and S. Szymczak, "Evaluation of variable selection methods for random forests and omics data sets," *Brief Bioinform*, vol. 20, no. 2, pp. 492–503, 2017, https://doi.org/10.1093/bib/bbx124.

[30] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata Journal*, vol. 20, no. 1, pp. 3–29, 2020, https://doi.org/10.1177/1536867X20909688.

[31] R. Couronné, P. Probst, and A. L. Boulesteix, "Random forest versus logistic regression: A large-scale benchmark experiment," *BMC Bioinformatics 19*, pp. 19–270, 2018, https://doi.org/10.1186/s12859-018-2264-5.

[32] T. M. Therneau and E. J. Atkinson, "An introduction to recursive partitioning using the RPART routines," *Mayo foundation*, pp. 1–60, 2022.

[33] N. J. Tierney, F. A. Harden, M. J. Harden, and K. L. Mengersen, "Using decision trees to understand structure in missing data," *BMJ Open*, vol. 5, no. 6, pp. 1–11, 2015, https://doi.org/10.1136/bmjopen-2014-007450.

[34] R. Wang, "AdaBoost for feature selection, classification and its relation with SVM, A Review," *Phys Procedia*, vol. 25, no. 2012, pp. 800–807, 2012, https://doi.org/10.1016/j.phpro.2012.03.160.

[35] P. Schober and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation," *Anesth Analg*, vol. 126, no. 5, pp. 1763–1768, 2018, https://doi.org/10.1213/ANE.0000000000002864.

[36] M. M. Mukaka, "Statistics corner: A guide to appropriate use of correlation coefficient in medical research.," *Malawi Med J*, vol. 24, no. 3, pp. 69–71, Sep. 2012, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23638278

[37] S. Kumar and I. Chong, "Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states," *Int J Environ Res Public Health*, pp. 2–24, 2018, https://doi.org/10.3390/ijerph15122907.

[38] S. Greenland *et al.*, "Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations," *Eur J Epidemiol*, vol. 31, no. 4, pp. 337–350, 2016, https://doi.org/10.1007/s10654-016-0149-3.

[39] L. Song, P. Langfelder, and S. Horvath, "Comparison of co-expression measures: Mutual information, correlation, and model based indices," *BMC Bioinformatics*, pp. 13–328, 2012, https://doi.org/10.1186/1471-2105-13-328.

[40] N. Barraza, S. Moro, M. Ferreyra, and A. de la Pena, "Mutual information and sensitivity analysis for feature selection in customer targeting: A comparative study," *J Inf Sci*, vol. 45, no. 1, pp. 53–67, 2019, https://doi.org/10.1177/0165551518770967.

[41] P. Laarne, M. A. Zaidan, and T. Nieminen, "ennemi: Non-linear correlation detection with mutual information," *SoftwareX*, vol. 14, pp. 2–5, 2021, https://doi.org/10.1016/j.softx.2021.100686.

[42] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural Comput Appl*, vol. 24, no. 1, pp. 175–186, 2014, https://doi.org/10.1007/s00521-013-1368-0.

[43] S. Akanmu and S. Jaja, "Knowledge Discovery in Database: A knowledge management strategic approach," Oct. 2012.

[44] H. Patel, D. S. Rajput, G. T. Reddy, C. Iwendi, K. A. Bashir, and O. Jo, "A review on classification of imbalanced data for wireless sensor networks," *Int J Distrib Sens Netw*, vol. 16, no. 4, pp. 1–15, 2020, https://doi.org/10.1177/1550147720916404.

[45] F. Afghah, A. Razi, R. Soroushmehr, H. Ghanbari, and K. Najarian, "Game theoretic approach for systematic feature selection: Application in false alarm detection in intensive care units," *Entropy*, vol. 20, pp. 1–16, 2018, https://doi.org/10.3390/e20030190.

[46] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 01–11, 2015, https://doi.org/10.5121/ijdkp.2015.5201.

[47] S. A. Hicks *et al.*, "On evaluation metrics for medical applications of artificial intelligence," *Sci Rep*, vol. 12, no. 1, pp. 1–9, 2022, https://doi.org/10.1038/s41598-022-09954-8.

[48] M. L. McHugh, "Interrater reliability:the kappa statistic," *Biochemia Medica (Zagreb)*, pp. 278–282, 2012, https://doi.org/10.11613/BM.2012.031.

[49] A. Venkatasubramaniam, J. Wolfson, N. Mitchell, T. Barnes, M. Jaka, and S. French, "Decision trees in epidemiological research," *Emerg Themes Epidemiol*, vol. 14, no. 1, pp. 1–12, 2017, https://doi.org/10.1186/s12982-017-0064-4.

[50] C. Porzelius, M. Schumacher, and H. Binder, "The benefit of data-based model complexity selection via prediction error curves in time-to-event data," *Comput Stat*, vol. **26**, pp. 293–302, 2011. https://doi.org/10.1007/s00180-011-0236-6.

[51] G. Stiglic, S. Kocbek, I. Pernek, and P. Kokol, "Comprehensive decision tree models in bioinformatics," *PLoS One*, vol. 7, no. 3, 2012, https://doi.org/10.1371/journal.pone.0033812.

[52] M. B. Kursa and W. R. Rudnicki, "Feature selection with the boruta package," *J Stat Softw*, vol. 36, no. 11, pp. 1–13, 2010, https://doi.org/10.18637/jss.v036.i11.

[53] K. Kang and J. Michalak, "Enhanced version of AdaBoostM1 with J48 Tree learning method," 1802.03522, Feb. 2018.

[54] G. Eibl and K. P. Pfeiffer, "How to make adaboost.m1 work for weak base classifiers by changing only one line of the code," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2430, pp. 72–83, 2002, https://doi.org/10.1007/3-540-36755-1_7.

[55] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

Frans Ramphele(MPHIL, University of Cape Town) is currently a Director in the Limpopo Department responsible for the Education Management Information System (EMIS). He has considerable knowledge and experience in the areas of ICT/IS management, software development, business intelligence, artificial intelligence and machine learning, data mining and database management Among other things, he is a member of the South African Higher Education Committee (HEDCOM) for e-Education, which oversees the development of ICT guidelines for e-Education.

Zenghui Wang *(PhD, Nankai University, China) is a National Research Foundation (NRF) C2-rated researcher in Electrical Engineering. He is currently a Professor at the University of South Africa (Unisa) in the Department of Electrical and Mining Engineering. Prof Wang specialises in Electrical Engineering and Computer Science. He is the leader of Intelligent System research group/laboratory.*

He has been involved in more than ten major research projects in South Africa and China. Up until February 2022, he has approximately 180 papers published or accepted, including 90 ISI master indexed journal papers.

Adedayo Yusuff *(D.Tech.) is a lecturer and a researcher in Electrical Engineering. He is currently a Professor at the University of South Africa (Unisa) in the Department of Electrical and Mining Engineering. Adedayo Yusuff has extensive experience in the industry in the areas of signal processing and control, network optimisation, and machine translation. Prof Yusuff specialises in adaptive electric power transmission networks; application of computational intelligence schemes for operation, control, and protection of advanced power grid; and integration of intermittent and renewable energy sources to power grid. He has been involved in research projects in South Africa and Nigeria.*