Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/170209

Predictive Analysis of Misuse of Alcohol and Drugs using Machine Learning Algorithms: The Case of using an Imbalanced Dataset from South Africa

Alexander Boateng [1], *Christopher Odoom* [1], *Eric Teye Mensah* [1], *Sarah Mensah Fobi* [1], *and Daniel Maposa* [2,*]

[1] Department of Statistics and Actuarial Science, Kwame Nkrumah University of Science and Technology, Ghana
[2] Department of Statistics and Operations Research, University of Limpopo, South Africa

Received: 2 Nov. 2022, Revised: 22 Dec. 2022, Accepted: 15 Feb. 2023
Published online: 1 Mar. 2023

Abstract: Drug and alcohol misuse has become a significant global distraction to the development of the youths and future leaders, especially in South Africa. The ideal solution to these unhealthy practices is for institutions, governments, and individuals to put preventive measures such as counselling, sanctions, and fines to help control drug and alcohol misuse. However, it is challenging to know the group of people or individuals engaged in such deteriorating activities by merely monitoring individuals physically. It makes it difficult even to plan the measures to control the issue. Therefore this paper has applied and compared six supervised machine learning algorithms to predict alcohol abuse and drugs across the nine provinces in South Africa to propose an ideal predictive model for detecting drug and alcohol misuse. The data used in the study was extracted from the 2019 General Household Survey conducted by Statistics South Africa, South Africa. The algorithms used in this paper include Random Forest, Naive Bayes, Support Vector Machines, Logistic Regression, Artificial Neural Networks, and Decision Tree. Results from the study identified Decision Tree to provide the highest recall of about 82.76 per cent, for alcohol abuse and drugs prediction, compared to the other five algorithms. In terms of the features and their importance, we found males across all the educational levels, mostly youth living in the Western Cape and Free State, to have played a vital role in the classification process. As part of the implication of the results in terms of policy formulation, we will edge the South African National Council on Alcohol and Drugs (SANCA) to draw up intervention programs to address issues of alcohol abuse and drugs, targeting all six attributes across all the provinces, especially Free State and North West provinces of South Africa. In conclusion, the approach used in this paper has effectively revealed an appropriate algorithm with a very high recall, reducing false negative, which means a successful reduction in misclassification of the important class. These results are reliable and valid for detecting drug and alcohol misuse at a low cost compared to other rigorous and demanding approaches.

Keywords: Alcohol; Drug abuse; Decision Tree; Naive Bayes; Random Forest; Supervised Machine Learning Algorithms; Support Vector Machine

1 Introduction

Drug addiction is a continuously worsening condition marked by compulsive use of addictive substances regardless of unfavourable effects on the people and community [1]. Alcohol and drug addiction are becoming a global trend affecting developed and developing nations. Alcoholism, drug addiction, and cigarette smoking have become major public health issues. Unable to regulate alcohol consumption can signify a more significant challenge, leading to two different issues, namely, alcohol abuse and alcoholism. According to the National Institute of Alcohol Abuse and Alcoholism [2], around 18 million persons in the United States suffer from alcohol-related disorders [2]. These illnesses can be both frustrating and dangerous. Abuse and addiction to alcohol can result in serious health problems. Alcohol, for example, exacerbates some diseases like osteoporosis and can even cause cancer. Other health problems, such as heart disease, are more difficult to diagnose by alcohol addiction.

In 2016, the abuse of alcohol accounted for 3 million (5.6 percent) deaths worldwide [3]. Alcohol abuse in relation to mortality rate is higher than the mortality rate caused by tuberculosis, HIV/AIDS, and diabetes [4]. In 2017, the abuse of alcohol and drugs accounted for 11.8

* Corresponding author e-mail: danmaposa@gmail.com / Daniel.maposa@ul.ac.za

million deaths globally [5]. The number of people aged between 16 to 64 who used drugs in 2016 was 30 per cent higher than those of the same age group who used drugs in 2009 [6]. In South Africa, 80 per cent of young male deaths related to alcohol and drug consumption is estimated to be two times more than the world norm [7]. Studies have also shown that Cannabis is the primary drug among South Africans younger than 20 years old [8]. Again, South African provinces dominant in drug abuse are Mpumalanga, Limpopo, and Western Cape. Whereas Free State, Northern Cape, North West, and Eastern Cape are dormant with alcohol abuse [8]. Therefore, we need to understand the dynamics and risk factors responsible for the trend and suggest remedies to the predominantly dominated provinces using drugs and abusing alcohol. It is for this reason that the study is being conducted.

Several studies have been conducted on alcohol abuse and drugs in South Africa (see, for example, [9], [10], [11], and [12]) with associated risk factors such as education level, employment status, race, marital status, peer pressure, and family background. In the area of methodology, most researchers including but not limited to [13], [14], [15], and [11] applied classical models such as regression analysis to examine risk variables associated with alcohol abuse and drugs. Regression analysis is suitable for evaluating a priori specified impacts, but it cannot capture indeterminate inter-relationships between components, thus limited. Machine learning (ML) techniques can overcome these drawbacks by employing a variety of probabilistic, optimisation, and statistical methods to uncover hidden and complicated relationships and patterns in data.

By applying ML to alcohol abuse and drug data, important decisions can be taken, and predictions can be made. The predictive analysis with ML aims to diagnose risk factors associated with alcohol abuse and drugs across all the nine provinces of South Africa with the best possible accuracy. Machine learning techniques are rapidly utilised to develop algorithms that have been demonstrated to be more reliable than traditional methods in terms of prediction ([16], [17]). Many medicine and public health disciplines have recently relied on machine learning algorithms (see, for example, [18], [19], [20]). However, the approach is yet to be implemented in this study using population-based data from Statistics South Africa. Indeed, we found a few studies that have used machine learning models to predict risks and had satisfactory predictive accuracy results in our literature search. [21], for example, used classification trees and random forests to estimate the vulnerability of ever-married women in India between the ages of 15 to 40 years of domestic abuse events. [22] also built an intimate partner violence perpetration triage tool that was applied to classify youth who are in danger of instigating intimate partner violence using supervised machine learning algorithms such as support vector machines and artificial neural networks, random forest, and logistic regression. This paper focuses on building a predictive model using six machine learning algorithms to investigate alcohol consumption and drug-related consequences within all the nine provinces in South Africa to foster the understanding of risk and protective factors associated with subjects.

Indeed, there are limited or no studies about machine learning methods in analysing alcohol abuse and drug-related issues in the South African context. This is the gap we intend to fill in this regard. We seek to contribute to the literature on ML by addressing the following specific objectives. Firstly, to construct predictive models with six ML algorithms that can efficiently classify alcohol abuse and drug based on some risk factors across the nine provinces of South Africa. Secondly, to discover major features associated with alcohol abuse and drug risk. We seek to build ML models, including Randon Forest (RF), Naive Bayes (NB), Support Vector Machines (SVM), Logistic Regression (LR) and Artificial Neural Networks (ANN) and Decision Tree (DT) using an imbalanced dataset, to predict the occurrence of alcohol abuse and drugs across all the provinces, given the risk factors. Models proposed in this study can efficiently detect subgroups of vulnerable people at high risk of abusing drugs and alcohol to increase intervention efficiency. The paper's outline: Materials and Methods are presented in Section 2. The results and discussion are presented in Section 3, followed by the conclusion and policy recommendations in Section 4.

2 Materials and Methods

This section describes; the sources of data, data preprocessing and collection, model building and model evaluation metrics. The algorithms utilised in the model-building process include Naive Bayes (NB), Logistic Regression (LR), Support Vector Machines (SVM), Randon Forest (RF), and Artificial Neural Networks (ANN) and Decision Tree (DT). The ROC metric, such as sensitivity, recall and precision, are applied to evaluate the model's performance in classifying alcohol abuse and drugs in the nine provinces of South Africa. The methodology used for this study is displayed in Figure 1 below.

Fig. 1: Steps in building the methodology for ML algorithm.

2.1 Data Sources, Collection and Preprocessing

The data for the analysis was secondary data obtained from the 2019 General Household Survey collated by Statistics South Africa. It comprises 48164 observations and six variables across all the nine provinces: Eastern Cape, Free State, Gauteng, Kwa Zulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West, and Western Cape. The data description is given in Table 1. Data preprocessing was done to remove outliers and features with missing values. The dataset was split into a 70 per cent training set and a 30 per cent test set using the repeated k-fold cross-validation and the slipt resampling techniques.

Table 1: Variable description and data set code.

Variable	Description	Data code
Alcohol or drug abuse	Binary response from the participants of a survey: yes or no	**Alc_drug**
Provinces	The nine provinces of South Africa: Eastern Cape (EC), Free State (FS), Gauteng (GP), Kwa-Zulu Natal (KZN), Limpopo (LP), Mpumalanga (MP), Northern Cape (NC), North West (NW), and Western Cape (WC).	***Province***
Gender	Gender of a participant: Female and male.	***Gender*** ***Race***
Age	The participants' age groups: Youth (12-24), Adults (25-64), and Elderly (65 years and over).	***Age***
Marital Status	The marital status of the participants: Divorced, Married, and Single.	***MaritalS***
Education-al level	The highest education level reached by the participants: Primary, Secondary, Tertiary, and other.	***EduLevel***

2.2 Model Building Process

2.2.1 Machine Learning

The use of machine learning (ML) algorithms to represent social, psychological, and biological phenomena is becoming more common ([16], [23], [24], [25], [26], [27]). Machine learning is a branch of computer science that explores how algorithms learn by themselves over time ([28], [29]). This subdiscipline, which lies at the crossroads of computer science and statistics, is called Artificial Intelligence (AI) ([28], [29]). Today, machine learning is an essential component of modern business and research. It employs algorithms and artificial neural network models to help computer systems gradually improve their performance. Machine learning algorithms create a mathematical model using sample data – also known as "training data" – to make decisions without being explicitly programmed.

2.2.2 Supervised and Unsupervised Algorithms

An algorithm, f a set of instructions, describes how a machine learns through experience. Learning, therefore, means that the algorithms have identified how to consistently predict an outcome (Y) given a set of predictor variables (X). The method uses the predictors $f(X)$ to predict a new outcome $\hat{Y}$ that is as close to the original or observed Y as possible (i.e.)

$$Y = f(x) + \varepsilon \tag{1}$$

where, ε is an error term and $f(X) = \hat{Y}$. This sort of learning is known as supervised learning since the algorithm is focused on the outcome Y. Unsupervised learning, on the other hand, is a process in which an algorithm is designed to summarise a (big) set of predictors (Z) to one (or a few) outcomes [28].

Naive Bayes Algorithm The Naive Bayes (NB) method is a supervised learning algorithm for addressing classification issues based on the Bayes theorem. Naive Bayes primarily uses in-text classification tasks requiring a large training dataset. The NB Classifier is a simple and effective classification method that aids in developing fast machine learning models capable of making quick predictions. NB is a probabilistic classifier that makes predictions based on an object's probability. Spam filtration, sentiment analysis, and article classification are common uses of the Naive Bayes Algorithm [32].

The Naive Bayes classification is determined by

$$p(Y = C_k|X_1,...,X_d) = \frac{p(Y = C_k)\prod_i p(X_j|Y = C_k)}{\sum_i p(Y = y_i)\prod_i p(X_j|Y = y_i)} \tag{2}$$

where, Y and X are random variables, respectively, denoting label and feature and C_k denoting possible K classes for $k \in [1,...,K]$. If we are interested only in the most probable value of Y, then we have the NB classification rule expressed in equation (1) as

$$Y \leftarrow \underset{C_k}{\arg\max} \frac{p(Y = C_k)\prod_i p(X_j|Y = C_k)}{\sum_i p(Y = y_i)\prod_i p(X_j|Y = y_i)} \tag{3}$$

Now, because the denominator does not depend on C_k. The value of the formulation above can be simplified to the equation (4)

$$Y \leftarrow \underset{C_k}{\arg\max} \left(p(Y = C_k) \prod_i p(X_j | Y = C_k) \right) \quad (4)$$

Logistic Regression Algorithm Logistic regression (LR) is another robust supervised machine learning algorithm used for binary classification problems when the target is categorical. Logistic regression uses a logistic function in equation (5) to model a binary output variable.

$$\log \left(\frac{y}{1-y} \right) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_d x_d \quad (5)$$

where, y is the probability of belonging to the class of interest, θ_i are the coefficients or the weights on the features and x_i's are the features. Replacing y with $p(X)$, we get the equation (6)

$$p(X) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_d x_d)}} \quad (6)$$

These probabilities are used to classify an observation as group A or B.

Support Vector Machine Algorithm The Support Vector Machine (SVM) is a popular supervised learning technique that solves classification and regression problems. However, it is mainly utilised in machine learning for classification problems. SVM algorithm's purpose is to find the optimum line or decision boundary for categorizing n-dimensional space into classes so that additional data points can be readily placed in the correct category in the future with the best decision boundary is called a hyperplane. The extreme points or vectors that assist create the hyperplane are chosen via SVM. Support vectors are extreme instances, and the algorithm is called a Support Vector Machine. Figure 2 below shows how a decision boundary or hyperplane is used to classify two different categories:

Random Forest Algorithm Random Forest (RF) is a well-known machine learning algorithm that uses the supervised learning method for classification and regression problems. Random forest is based on ensemble learning, which integrates several classifiers to solve a complex issue and increase the model's performance. Random Forest is a classifier with several decision trees on various datasets' subsets. According to the name RF, it takes the average to enhance the predicted accuracy of that dataset (see Figure 3). Instead of relying on a single decision tree, the random forest collects the forecasts from several trees and predicts the final output based on the majority votes of predictions.

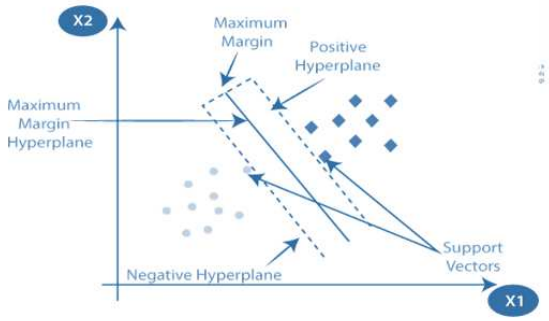


Fig. 2: Support Vector Machine structure (source: www.javatpoint.com).

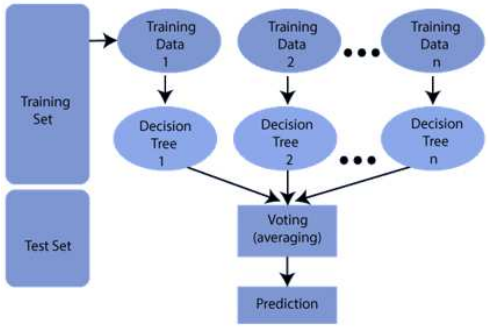


Fig. 3: Random forest structure (source: www.javatpoint.com).

Since the random forest combines numerous trees to forecast the dataset's class, some decision trees may correctly predict the output while others may not. However, a good result is expected when all trees are combined. As a result, two assumptions for a better Random Forest classifier are as follows: Firstly, the dataset's feature variable should have some actual values so that the classifier can predict accurate results rather than guesses. Secondly, each tree's predictions must have very low correlations. For instance, assume you have a dataset with various fruit photos. As a result, the Random Forest classifier is given this dataset. Each decision tree is given a portion of the dataset to work with. During the training phase, each decision tree generates a prediction result. The Random Forest classifier predicts the final decision based on most outcomes when a new data point appears. As it can be envisaged in Figure 4 below: Indeed, equation (7) is used for regression purposes. This formula computes the distance between each node and the predicted actual value, assisting you in determining which branch is the best choice for your forest.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (f_i - y_i)^2 \quad (7)$$

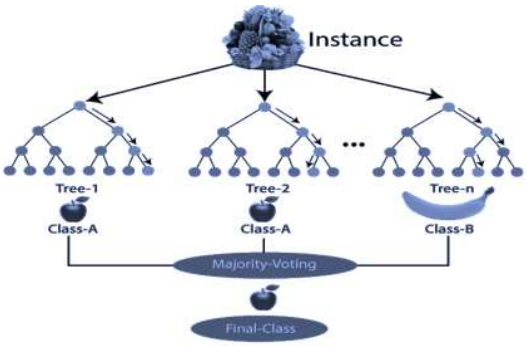


Fig. 4: Random forest structure (source: www.javatpoint.com).

where, N is the number of data points, f_i is the value returned by the model and y_i is the actual value for the data points i.

The random forest uses the Gini index to identify which class an observation belongs to for classification purposes, as indicated in equation (8). Equation (8) calculates the Gini of each branch on a node based on type and probability, deciding which branch is more likely to occur.

$$\text{Gini} = 1 - \sum_{i=1}^{C} (P_i)^2 \tag{8}$$

where, P_i represents the relative frequency of the class you are observing in the dataset C denotes the number of classes.

Artificial Neural Network Algorithm A computational model was constructed based on the human brain. Voice recognition, picture identification, and robotics using Artificial Neural Networks are just a few recent developments in Artificial Intelligence. Artificial Neural Networks (ANNs) are computer simulations that are biologically inspired and execute specific tasks. An artificial neural network (ANN) is a biologically inspired network of artificial neurons configured to perform specific tasks, known as an artificial neural network (ANN). An input layer, hidden layer, and output layer are the three layers of an ANN. The Artificial Neural Networks (ANN) method takes in data and calculates the weighted total of the data, including bias. This work makes use of the transfer function, which is listed below.

$$\text{Transfer function} = \sum_{i=1}^{n} W_i \times X_i + b \tag{9}$$

where, W_i denotes the weight given to ith the input X_i is the ith input and b is a constant.

The sigmoidal hyperbolic is used to approximate output from net inputs, and it is given by

$$\hat{y} = \sigma(x) = \frac{1}{1 + e^{-\nu x}} \tag{10}$$

where, $\hat{y}$ denotes the sigmoid activation function and ν the steepness parameter, and the output obtained after forward prorogation is the predicted value.

Decision Tree Unlike other supervised learning algorithms, the decision tree algorithm can also solve regression and classification problems. The purpose of using a decision tree is to create a training model for predicting the class or value of a target attribute by learning simple decision rules from previous data (training data). In Decision Trees, we begin at the tree's root to predict a class label for a record. The results of the root attribute are compared to the values of the record's attribute. We follow the branch corresponding to that value and proceed to the next node based on the comparison.

2.3 Evaluation Measures for the Algorithms

Measurement of performance is critical in Machine Learning. The evaluation of algorithms is the final step of the prediction model. This study evaluates the prediction results using various evaluation metrics like classification accuracy, confusion matrix, kappa statistics F1-score, and ROC metrics.

2.3.1 Receiver Operating Characteristic (ROC)

In Machine Learning, performance measurement is an essential task, and when it comes to a classification problem, we can count on a ROC Curve. ROC curve is a graphical representation of a binary classifier system's diagnostic performance when modified discrimination threshold.

2.3.2 Accuracy Measures

Accuracy measure is the ratio of correct predictions to the total number of input samples. This is given by

$$\text{Accuracy} = \frac{\text{total number of correct predictions}}{\text{total number of predictions made}} \tag{11}$$

2.3.3 Confusion Matrix

Confusion matrix gives us a matrix as output and describes the complete performance of the model (see Table 2).

Table 2: Confusion matrix.

Predicted values \ Actual Values	Alc_drug (Yes)	Alc_drug (No)
Alc_drug (Yes)	TP	FP
Alc_drug (No)	FN	TN

where, TP denotes True Positive, FP denotes False Positive, FN denotes False Negative, and TN denoting True Negative. Accuracy can also be calculated from the confusion matrix by taking an average of the values across the main diagonal. It is given as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{12}$$

2.3.4 Sensitivity/Recall

Sensitivity determines the proportion of the positive class correctly classified, and it expressed in equation

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

2.3.5 Specificity

Specificity refers to the proportion of negative cases class correctly classified (see equation (13))

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{14}$$

2.3.6 Precision

Precision measures the probability of a sample classified as positive as actually positive. It is illustrated in equation (14)

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

2.3.7 F1-Measure

F1-Measure assesses the precision of a test. F1-Measure represents the harmonic Mean of accuracy and Recall. The F1 measure has a range of $[0,1]$. It informs you of your classifier's precision as well as its robustness. Mathematically, it is given as

$$\text{F1-Measure} = \frac{1}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}} \tag{16}$$

3 Results Presentation

In this study, six supervised machine learning algorithms are applied and compared. These include Randon Forest (RF), Naive Bayes (NB), Support Vector Machines (SVM), Logistic Regression (LR) and Artificial Neural Networks (ANN) and Decision Tree (DT). All the algorithms were applied to the alcohol abuse and drug dataset from South Africa. The dataset was divided into 70 per cent training and 30 per cent datasets using the split and repeated k-fold cross-validation sampling techniques. The dataset is imbalanced, with more observations in the negative (99.79 per cent) and few in the Positive class (0.21 per cent) - this imbalanced nature of the dataset is shown in Figure 5 below. Supervised

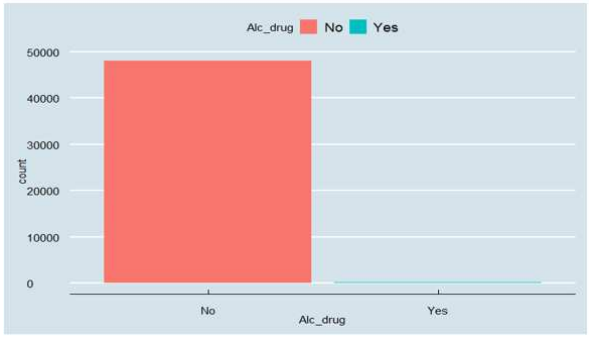


Fig. 5: Alcohol abuse and drugs.

algorithms are known to learn very badly from an imbalanced dataset. So to avoid making poor predictive models, this study utilized the combination of oversampling and undersampling methods, termed "both". Oversampling and undersampling ensure that the training dataset has enough observations in both classes. Oversampling is when the minority class increases or makes the same number as the majority class. This may be done by duplicating the minority class to reach the number of the majority class. Undersampling is when the majority of the class is reduced to the same number as the minority class. Both techniques combine oversampling and undersampling methods, which were used to generate a more balanced dataset to train the six algorithms.

All the confusion matrices for the six algorithms are shown in Table 3 below:

Table 4 below shows model accuracy measures for the comparison. The Naive Bayes algorithm recorded the highest accuracy of 94.07%, followed by Artificial Neural Networks (72.94%), Logistic Regression (72.30%), Support Vector Machine (71.69%), Random Forest (71.12%) and Decision tree (57.60%). This accuracy measure is unsuitable for comparing these algorithms since the test set is imbalanced.

Table 3: Confusion Matrix.

Panel	Predicted	Actual	
		Yes	No
A	Yes	24	6165
	No	5	8358
B	Yes	21	4194
	No	8	10329
C	Yes	22	4113
	No	7	10410
D	Yes	9	843
	No	20	13680
E	Yes	22	3931
	No	7	10592
F	Yes	21	4023
	No	8	10500

Note: A = DT; B = RF; C = SVM; D = NB; E = ANN; and F = LR

Table 4: Algorithm Performance Measures I.

Algorithm	Accuracy	Kappa	F1-score	Precision
A	0.5760	0.0038	0.0077	0.0038
B	0.7112	0.0060	0.0099	0.0050
C	0.7169	0.0066	0.0106	0.0053
D	0.9407	0.0166	0.0204	0.0106
E	0.7294	0.0071	0.0110	0.0056
F	0.7230	0.0064	0.0103	0.0052

Note: A = DT; B = RF; C = SVM; D = NB; E = ANN; and F = LR

Table 5 below gives the main evaluation measures employed in this study to compare the algorithms. These measures are recall/sensitivity and specificity. These measures are employed because predicting actual alcohol abuse and the drug is of great interest to organisations that manage drug abuse. Many interviewees hide their drug abuse status in this study, so getting a model that predicts more true positives as positive is essential in designing interventions. The decision tree recorded the highest sensitivity of 82.79%, followed by SVM(75.86%) and ANN(75.86%). Interestingly, the Naive Bayes classifier recorded the worst recall (31.03%). This observation confirms why accuracy may not be a good measure for an imbalanced dataset. Based on the findings of this paper, we conclude that the Decision Tree (DT) is the most suitable model for predicting alcohol abuse and drugs since it can fish out the highest percentage of positive classes compared to other algorithms.

Table 5: Algorithm Performance Measures II.

Algorithm	Recall	Specificity
A	0.8276	0.5755
B	0.7241	0.7112
C	0.7586	0.7168
D	0.3103	0.9420
E	0.7586	0.7293
F	0.7241	0.7230

Note: A = DT; B = RF; C = SVM; D = NB; E = ANN; and F = LR

Figure 6 shows the features importance plot for the algorithm with the best recall. The features with the longest bar are good predictors of alcohol abuse and drugs.

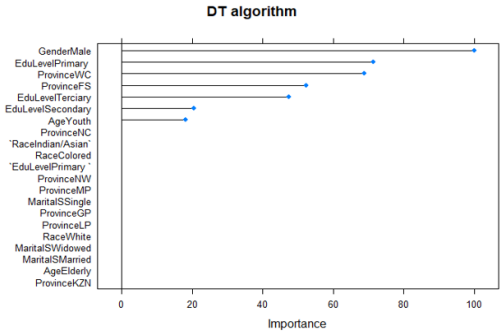


Fig. 6: Feature importance for the algorithm with the highest recall.

4 Discussion

The research found that "both" version of the resampling techniques, which combines both undersampling and oversampling, is more appropriate for resolving the imbalanced nature of the dataset compared to the individual methods. Most existing papers employed only one technique to resolve the imbalance dataset problem. However, these approaches are less effective than the combination of both methods, which is illustrated in this paper.

In addition, the research found out that the decision tree algorithm coupled with the "both" technique is appropriate for building a predictive model for the dataset. This is because of the higher sensitivity of the decision tree algorithm compared with other algorithms. This result is interesting since most algorithms such as

Random forest, Naive Bayes and support vector machines perform better than a decision tree. However, this result shows otherwise, implying that a simple decision tree is enough in some situations. Building a huge decision tree such as a random forest is not always required in all instances, such as a more sensitive issue like the one demonstrated in this paper.

The study also found that to build only an accurate predictive model without considering the negative effect (its inability to predict more yes for drug and alcohol abuse). The Naive Bayes algorithm is appropriate for such an aim. This is because the Naive Bayes algorithm recorded an accuracy measure of 94.07 per cent, which is very high. This result is not very relevant to this study but very worthy noting. In most predictive modelling problems, the model with the highest accuracy is preferred over others. However, in the situation described in the paper, going via this approach will only produce an algorithm that will rather fail to predict more positives. Although the Naive Bayes algorithm gave the highest accuracy, the false negative rate is very high, which is a significant concern to this paper. This result also implies that, in situations like the one in this paper, it is vital to consider other performance measures such as recall, specificity, false negative and false positive but not accuracy because accuracy is a bad measure of performance for an imbalanced dataset where the positive class is sensitive. Sensitive cases are issues that people are very good at hiding from the public, and it isn't easy to make them known to anyone. The nature of drug and alcohol misuse makes it worth predicting without the individual's asking. However, the concentration must be on the number or percentage of false negatives produced by the model.

The research also found out that males, citizens with primary education, province WC, and province FS are the critical attributes used in classifying drug and alcohol misuse in South Africa by the decision tree algorithm. This paper also provides some key attributes that policymakers must pay attention to when planning to control drug and alcohol abuse in South Africa. People with only primary education seem to contribute immensely to drug and alcohol abuse in South Africa. This may be due to their inability to understand the side effect on their body and dignity.

Finally, the research also found an interesting result about the Naive Bayes algorithm used for situations like ours. It recorded the highest accuracy but unfortunately had a very low recall of about 31 per cent. This finding tells researchers to focus on their goal rather than constantly seeking an accurate predictive model. Moreover, this paper used census data which is not always the case for imbalanced dataset research in literature. It also used a dataset made up of only categorical attributes, which differs from most research on similar work. Most researchers used only numeric attributes or a combination of both numeric and categorical. This paper also used standard or popularly used algorithms that have been improved but still effectively produced very efficient models like the one we have in this paper.

Nevertheless, this research is limited to some imbalanced dataset correction techniques because of the categorical nature of the attributes. Other resampling techniques, such as the synthetic minority oversampling method (SMOTE), are known to produce better samples for building models but are rarely used where the attributes are categorical, as addressed in this paper. The SMOTE method is an effective resampling technique to address the class imbalance [31].

5 Conclusions, Recommendations and Policy Implications

Predictive analysis of alcohol abuse and drug use can alter how medical researchers, particularly psychiatrists, gain insights from medical data and make decisions. We used six popular machine learning algorithms for predictive analysis in this paper. Random Forest (RF), Naive Bayes (NB), Support Vector Machines (SVM), Logistic Regression (LR) and Artificial Neural Networks (ANN) and Decision Tree (DT) algorithms were applied in the study. Predictions about alcohol abuse and drug were made using an alcohol and drug abuse dataset with 48168 records and six attributes. A simple decision tree is appropriate for predicting alcohol and drug misuse, which is revealed by the research. This result means that, despite the inclusion of other algorithms believed to perform better than decision trees in many instances, it does not render the effectiveness of a simple decision tree useless. The random forest, an aggregation of many simple decision trees, had a higher false negative than the simple decision tree. Unlike classical and bayesian regression, where predictions are made without identifying beneficial risk factors, this paper's model revealed that attributes that affect or dictate drug and alcohol misuse are gender, educational level, age group and province. This result is beneficial since its fills the gap identified in this research. It shows how efficient is machine learning algorithms in building predictive models over standard statistical models.

Additionally, the result was not based on any distributional assumption compared to classical models. In addition, the problem posed by imbalanced datasets on classification algorithms is severe; it always leads the algorithm to pay attention to the majority class, even though the minority class is of interest to researchers. This paper proposed a resampling method which combines oversampling and undersampling to create balanced classes for the model building. Though these methods both have their advantages which is to create a new sample with some level of balance for the classes, however, their disadvantage is the fact that these methods are skewed to one side, either by sampling the minority class to match the majority or sampling the majority to

check the minority class. This skewed nature is prevented by the algorithm or technique used in this paper, which combines both methods. This is an essential finding for model builders with a dataset made up of only categorical variables. In terms of the features and their importance, we found males across all the educational levels, mostly youth living in the Western Cape and Free State, to have played a vital role in the classification process. This implies that special attention should be given to these features to develop interventions for alcohol abuse and drugs. As part of the implication of the results in terms of policy formulation, we will edge the South African National Council on Alcohol and Drugs (SANCA) and Anti-Substance Abuse Program of Action, among others, to draw up intervention programs aiming to address issues of alcohol abuse and drugs targeting almost all the six attributes across all the provinces, especially Free State and North West provinces of South Africa. As part of the implication of the results in terms of policy formulation, we will edge the South African National Council on Alcohol and Drugs (SANCA), Anti-Substance Abuse Program of Action to adopt some of these machine learning algorithms to draw up intervention programs aimed at diagnosing and addressing issues of alcohol abuse and drugs in South Africa.

Further research may consider comparing model sensitivities, false positives and false negatives across various techniques for dealing with imbalanced datasets.

Acknowledgement

The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] Koob, G. F., & Volkow, N. D. (2010). Neurocircuitry of addiction. Neuropsycopharmacology 35: 217?238.

[2] National Institute of Alcohol Abuse and Alcoholism (NIAAA): https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets /understanding-alcohol-use-disorder.

[3] World Health Organization, 2018. Global status report on alcohol and health 2018: executive summary (No. WHO/MSD/MSB/18.2). World Health Organization.

[4] Laslett, A. M., Room, R., Waleewong, O., Stanesby, O., Callinan, S., & World Health Organization. (2019). Harm to others from drinking: Patterns in nine societies. World Health Organization.

[5] Ritchie, H., & Roser, M. (2019). Drug use. Our World in Data.

[6] United Nations Office on Drugs and Crime, 2018. Drugs and age: Drugs and associated issues among young people and older people.

[7] Parker, B. 2019. 80% of SA's male youth deaths are alcohol-related and drug consumption is twice the world norm. Available: https://www.parent24.com/Family/Health/80-of-sas-male-youth-deaths-are-alcohol-related-and-drug-consumption-is-twice-the-world-norm-20180626 (Accessed 19 August 2021).

[8] Dada, S., Harker Burnhams, N., Erasmus, J., Lucas Charles Parry, W., Bhana Sandra Pretorius, A., & Weimann Helen Keen, R. (2018). Monitoring alcohol, tobacco and other drug abuse treatment admissions in South Africa. SACENDU (South African Community Epidemiology Network on Drug Use) April, 1-72.

[9] Branstrom, R., & Andreasson, S. (2008). Regional differences in alcohol consumption, alcohol addiction and drug use among Swedish adults. Scandinavian journal of public health, 36(5), 493-503.

[10] Mafa, P., Makhubele, J. C., Ananias, J. A., Chilwalo, B. N., Matlakala, F. K., Rapholo, S. F., ... & Freeman, R. J. (2019). Alcohol consumption patterns: A gender comparative study among high school youth in South Africa. Global Journal of Health Science, 11(2), 92-101.

[11] Shmulewitz, D., & Hasin, D. S. (2019). Risk factors for alcohol use among pregnant women, ages 15?44, in the United States, 2002 to 2017. Preventive medicine, 124, 75-83.

[12] Mbandlwa, Z., & Dorasamy, N. (2020). The impact of substance abuse in South Africa: a case of informal settlement communities. Journal of Critical Reviews; Vol. 7, Issue 19.

[13] Lukasiewicz, M., Falissard, B., Michel, L., Neveu, X., Reynaud, M., & Gasquet, I. (2007). Prevalence and factors associated with alcohol and drug-related disorders in prison: a French national study. Substance abuse treatment, prevention, and policy, 2(1), 1-10.

[14] Kenna, G. A., & Lewis, D. C. (2008). Risk factors for alcohol and other drug use by healthcare professionals. Substance Abuse Treatment, Prevention, and Policy, 3(1), 1-8.

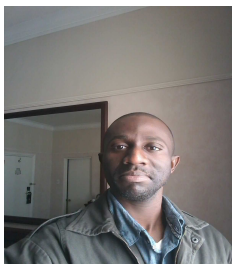
[15] Vythilingum, B., Stein, D. J., Roos, A., Faure, S. C., & Geerts, L. (2012). Risk factors for substance use in pregnant women in South Africa. South African Medical Journal, 102(11), 851-854.

[16] Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is machine learning? A primer for the epidemiologist. *American journal of epidemiology, 188(12)*, 2222-2239.

[17] Uddin, S., Khan, A., Hossain, M., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, 19, Article 281. https://doi.org/10.1186/s12911-019-1004-8

[18] Galatzer-Levy, I. R., Ma, S., Statnikov, A., Yehuda, R., & Shalev, A. Y. (2017). Utilisation of machine learning for prediction of post-traumatic stress: a re-examination of cortisol in the prediction and pathways to non-remitting PTSD. Translational psychiatry, 7(3),

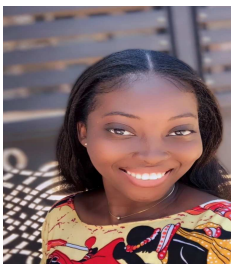
e1070-e1070.
[19] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.
[20] Yang, F., Wang, H. Z., Mi, H., & Cai, W. W. (2009). Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. BMC bioinformatics, 10(1), 1-14.
[21] Ghosh, D. (2007). Predicting vulnerability of Indian women to domestic violence incidents. Research and Practice in Social Sciences, 3(1), 48-72.
[22] Petering, R., Um, M. Y., Fard, N. A., Tavabi, N., Kumari, R., & Gilani, S. N. (2018). Artificial intelligence to predict intimate partner violence perpetration. In M. Tambe & E. Rice (Eds.), Artificial intelligence and social work (p. 195). Cambridge University Press.
[23] Luo, W., et al. (2016). Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *Journal of Medical Internet Research, 18(12)*, e323.
[24] Molina, M., & Garip, F. (2019). Machine learning for sociology. *Annual Review of Sociology, 45*, 27?45. https://doi.org/10.1146/annurev-soc-073117-041106
[25] Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *The Journal of Economic Perspectives, 31(2)*, 87?106.
[26] Wiemken, T. L., & Kelley, R. R. (2020). Machine learning in epidemiology and health outcomes research. *Annual Review of Public Health, 41*, 21?36.
[27] Wilkerson, J., & Casas, A. (2017). Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science, 20*, 529?544.
[28] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer Science & Business Media.
[29] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255?260.
[30] Aggarwal, C. C., & Clustering, C. R. D. (2014). Algorithms and applications.
[31] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.
[32] Aggarwal, C. C., & Clustering, C. R. D. (2014). Algorithms and applications.



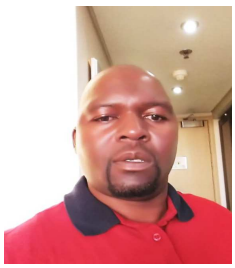
Alexander Boateng holds a PhD in Statistics from the University of Limpopo (formerly the University of the North) in South Africa. He is currently a lecturer in the Department of Statistics and Actuarial Science at Kwame Nkrumah University of Science and Technology (KNUST), Ghana. His research interests include statistical modelling and simulation, stochastic modelling, machine learning, time series analysis and generalised linear modelling. He is an external examiner for North-West University, South Africa. He has assessed several honours, master's, and doctoral theses from North-West University, the University of Johannesburg and the University of KwaZulu-Natal, all in South Africa and Kwame Nkrumah University of Science and Technology (KNUST), Ghana. He is a member of the South African Statistical Association (SASA), International Biometric Society (IBS) and International Statistics Institute (ISI). He has supervised several graduate and post-graduate students and has published several articles.

Christopher Odoom is a graduate student in the Department of Mathematics and Statistics, University of Massachusetts Amherst (UMASS), Amherst, USA. His research interests include Machine learning with a concentration on supervised machine learning algorithms, regression modelling, cluster analysis, multivariate data analysis, and epidemiology-related topics around diabetes, heart diseases and others.

Eric Teye Mensah received his BSc Statistics degree from the Kwame Nkrumah University of Science and Technology (KNUST), Ghana, in 2018. Currently, he is a Graduate Assistant in the Department of Statistics and Actuarial Science at KNUST. Eric Teye Mensah is passionate about the value of statistical thinking and statistical techniques to improve decision-making. He is interested in the following research areas: statistical modelling and simulation, machine learning, stochastic modelling, time series analysis and forecasting, and Generalized linear modelling.



Sarah Mensah Fobi is a graduate student in the Department of Mathematical Science, Montana State University (MSU), Bozeman, Montana, USA. Her research interests include regression analysis, time series analysis, statistics and data science education, and machine learning with a concentration in supervised machine learning algorithms.



Daniel Maposa is an Associate Professor in the Department of Statistics and Operations Research, School of Mathematical and Computer Sciences, Faculty of Science and Agriculture, University of Limpopo, South Africa. He holds a PhD degree in Extreme Value Theory (EVT) Statistics, a Master of Science degree in Operations Research and Statistics, and an Honours degree in Applied Mathematics. He is a National Research Foundation (NRF) C2-rated researcher in EVT statistics. He has published more than 35 journal research articles in internationally accredited journals, two book chapters and three conference proceedings. In his research activities, he has been all over the world attending international conferences and presenting his research work in statistics of extremes in countries such as New Zealand, Australia, China, Switzerland, Brazil, Morocco, Botswana and Malaysia. Daniel Maposa is a registered professional natural scientist (Pr.Sci.Nat.) in Statistical Sciences and Mathematical Sciences and is a member of International Statistical Institute (ISI) and a member of South African Statistical Association (SASA). He also regularly attends the South African Statistical Association (SASA) conference annually and the ISI World Statistics Congress organised bi-annually. He has supervised several Masters and Doctoral postgraduate students.