American Journal of Theoretical and Applied Statistics
2023; 12(5): 120-128
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20231205.14
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Classification of Contraceptive Use Among Undergraduate Students Using a Supervised Machine Learning Technique

Sammy Kiprop, Charity Wamwea, Herbert Imboga, Joel Chelule

Department of Statistics and Actuarial Science, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Email address:
kipropsammy61@gmail.com (Sammy Kiprop)

To cite this article:
Sammy Kiprop, Charity Wamwea, Herbert Imboga, Joel Chelule. Classification of Contraceptive Use Among Undergraduate Students Using a Supervised Machine Learning Technique. *American Journal of Theoretical and Applied Statistics*. Vol. 12, No. 5, 2023, pp. 120-128.
doi: 10.11648/j.ajtas.20231205.14

Received: September 26, 2023; **Accepted**: October 12, 2023; **Published**: October 28, 2023

Abstract: The Kenyan government in partnership with other stakeholders involved in providing family planning services have put in place various strategies and policies to increase uptake of contraceptives. This results in an increase in contraceptive prevalence rate (CPR), reduction of both total fertility rate (TFR) and sexually transmitted infections (STIs). Despite the various strategies and policies, the total fertility rate still remains high, while CPR has been unattained, respectively. The aim of this study was to classify contraceptives use among undergraduate students using a supervised machine learning technique. The target population constituted students at Jomo Kenyatta University of Agriculture and Technology (JKUAT) (Eldoret Campus). The study applied simple random sampling technique to obtain data from a sample of 252 using structured questionnaires. A decision tree classifier based on CHAID and C5.0 algorithms were used for classification. Pearson Chi-Squared statistic was used as feature selection technique to rank significant factors influencing contraceptives use based on their Chi scores. The findings show that the use of Chi-Squared feature selection led to contraceptives factors that were ranked higher having higher classification performance. The fitted decision tree model based on CHAID algorithm had a higher classification accuracy of 64.68% with 195 correct classifications as compared to the C5.0 decision tree model with accuracy of 61.18% with 163 correct classifications. The study findings contribute to a better insight on the classifications of contraceptives use among undergraduate students in Kenya. Hence, the government of Kenya can implement policies to enhance contraceptives awareness.

Keywords: Chi-Squared Feature Selection, Decision Tree Classifier, CHAID Algorithm, C5.0 Algorithm, Contraceptive's Use

1. Chapter One: Introduction

1.1. Background to the Study

Currently, the prosperity and independence of women, their families, and their communities depend on family planning. According to the United Nations Population Fund report UNFP, [19] the increased use of contraceptives has decreased maternal mortality by 30% globally. However, the associated side effects of using contraceptives include reduction of unintended pregnancies, spacing out pregnancies, and fewer high-risk pregnancies Darroch, J. E., & Singh, S., Ahmed, *et al.*, Brunner, H. *et al.* [5, 1, 2]. Noted that young girls who become pregnant unexpectedly face serious physical and mental health problems that place a heavy societal burden on them. In terms of the negative effects, Darroch, J. E. *et al*. [6] noted that the use of modern contraceptive methods has been shown to be the most effective means for young people to prevent unwanted pregnancies and complications arising from their use. There are different types of contraceptives, including hormonal and non-hormonal methods and their effectiveness and side effects kvary. Accurately classifying contraceptives is an essential task for healthcare providers and policymakers. Additionally, contraceptives play a crucial role in family planning, allowing couples to make informed decisions about the timing and spacing of their children Chiang, *et al*. [3].

Prior research has found a number of internal characteristics, including awareness, attitudes, and self-experiences, to be significantly correlated with the use of

contraceptives. In addition, various external factors are being confirmed as having an impact on young people's sexual health behaviors in comparison to the internal ones. Nevertheless, young people's usage of contraceptives has reportedly been linked to their socioeconomic level, sex education, family history, and access to contraceptives). The use of contraception by young people has also been linked to mental health histories and lifestyles children Sun, *et al*., Jiang, W., & Ha, L. [17, 11]. With the increasing availability and variety of contraceptive methods, it can be challenging for healthcare providers to choose the most appropriate contraceptive for their patients based on their individual needs and preferences.

The rate of contraceptive uptake with youth majorly in universities is still low. The findings by Kithuka. [32] showed that the use of condoms in JKUAT is still low and uptake of other contraceptives is not known. The descriptive statistics findings showed that out of 436 students in JKUAT subjected to the study on factors influencing contraceptive uptake amongst students, 55 percent (240) had experienced sexual intercourse while the use of contraceptives was still low at 34.2 percent (149). Therefore, as noted by Njoroge. [30] contraceptive costs, providers' attitude and students' contraceptive attitude and practice play great role in contraceptive uptake.

1.2. Objectives of the Study

The study focused to classify the use of contraceptives among undergraduate students using a Supervised Machine Learning Model. The specific objectives of the study were as follows:

1) To perform feature selection on the independent variables in relation to factors influencing contraceptives use.
2) To fit decision tree model on the contraception data based on CHAID and C5.0 algorithms.
3) To determine the performance of the fitted model
4) To determine the goodness of fit of the fitted model.

2. Chapter 2: Literature Review

2.1. Introduction

This section presents literature review in relation to past studies done on the classification of contraceptive use using supervised machine learning techniques. This section presents a review of feature selection and supervised machine learning techniques. In addition, this section presents the identification of research gaps for this study.

2.2. Feature Selection

A, Rehman. *et al.* [29] conducted feature selection, using Chi-Squared (CHI2), Information Gain (IG), and Least Absolute Shrinkage and Selection Operation (LASSO) to rank the relative importance of the relevant miRNAs. The findings showed that the miRNAs ranked higher by Chi-Squared method had higher classifier performance. Similar findings were found by Kuang., & Davison. [26] who propose learning weights based on the words' relative importance in the classification task using chi-square weights. Their findings depicted that embedding learned from the proposed algorithms outperform existing techniques by a relative accuracy improvement of over 9%.

Moreover Jin, *et al*. [25] used Chi-square for tag/gene selection for classification of cancers based on SAGE data using popular machine learning techniques such as Naive Bayes, SVM, Nearest Neighbor, RIPPER and C4.5. The findings suggest that Chi-square based gene selection can improve the performance of all the five classifiers investigated.

2.3. Overview Supervised Machine Learning Techniques Used Previously in the Classification of Contraceptives Usage

2.3.1. Support Vector Machine (SVM) Model

Ftye, *et al*. [8] Investigated the likelihood of contraceptive method usage among women in Ethiopia. This was based on 15,683 records of EDHS data where different machines learning algorithms were applied including support vector machine, Naive Bayes classifier, Neural Network and J48 Decision Tree classifier. The findings show that the support vector machine method achieved a higher classification accuracy of 86.17% as compared to other models.

Haq, *et al*. [9] used a support vector machine (SVM) algorithm to classify the contraceptive methods used by women in Bangladesh. The dataset used was collected through a survey of married women of reproductive age. The results showed that the proposed SVM model achieved an accuracy of 92.75%, indicating that it can be an effective tool for classifying contraceptive methods. SVM algorithm can be used to classify contraceptive methods used by women in Bangladesh.

2.3.2. Artificial Neural Network (ANN)

Haq, *et al*. [9] used a hierarchical logistic regression classifier based on logistic regression, random forest, least absolute shrinkage and selection operation (LASSO), Naives Bayes, neural network, AdaBoost and classification trees to predict contraceptive practice among women aged 15-49 years in Bangladesh. The findings suggest that the neural network method achieved the highest accuracy of 79.34% with AUC of 86.90%. Furthermore, the Cohen's Kappa statistic based on neural network was 0.5626 showing the most extreme discriminative ability.

2.3.3. Decision Tree Classifier

Tangirala, S. [18] used decision tree classifier algorithm to evaluate the goodness of the split based on GINI index and Information gain individually. The findings show that the classification built by applying the two splitting indices achieved the same accuracy of 78.5%. Pandya, R., & Pandya, J. [13] compared the performance of ID3, C4.5 and C5.0 decision tree algorithms to improve decision tree with feature selection and reduced error pruning. The findings show that

C5.0 algorithm achieved a higher accuracy and more efficient results as compared to ID3 and C4.5 algorithm.

Ibad, *et al*. [10] used decision tree classifier in the classification of information related to contraceptives use in BKKBN. The findings show that the Method Information Index has the highest influence on the selection of contraceptives, followed by the variables of domicile, ownership of children and place of residence. The decision tree model has an accuracy value of 88.99% hence can be used to predict the model in the population.

2.3.4. Hybrid Model

Haq, *et al*. [9] used machine learning algorithms for predicting contraceptive use among women in Nigeria. The study aimed to predict contraceptive use among women in Nigeria using supervised machine learning algorithms. Logistic regression, decision tree, and random forest algorithms were used to classify the contraceptive use status of women based on demographic and socioeconomic factors. The results showed that the decision tree algorithm had the highest accuracy of 80.7%.

3. Chapter 3: Methodology

3.1. Research Philosophy and Design

This study employed the use of a positivist epistemological approach and a quantitative research design based on a survey strategy to classify contraceptive use among undergraduate students using supervised machine learning technique. The use of a positivist approach allows the researcher to analyze the research objectives, before putting it to the test with data and facts Collis, J. & Hussey, R. [4]

3.2. Population and Sampling Criteria

The targeted population used in this study were undergraduate students from JKUAT University. Therefore, a random sampling technique was applied in this study where undergraduate students at JKUAT University (Eldoret Campus) were requested to participate. Based on this sampling technique, undergraduate students from other branches of JKUAT University were omitted as Eldoret Campus is where I am located and more accessible. This ensured that a representative sample was chosen consisting undergraduate students. Based on this study, a total sample of 255 undergraduate students were selected using simple random sampling technique as follows. First, the sample size was determined by the Fisher's formula below, assuming that p=0.5 and 95% confidence interval.

$$\mathrm{n} = \frac{\left(\frac{z\alpha}{2}\right)^2 \mathrm{pq}}{\mathrm{d}^2} \tag{1}$$

$$= \frac{(1.96)^2 * 0.5 * 0.5}{0.05^2} = 384.16 = 385\ students$$

However, the total population for the University was less than 10000 for the previous semester, hence the finite correction factor was employed to determine the correct sample size that was used in this research. This reduced the sample size as shown below.

$$n = \frac{n_\alpha N}{n_\alpha + (N-1)} \tag{2}$$

$$= \frac{385 * 750}{385 + 749} = \frac{288{,}750}{1{,}134} = 254.63 = 255\ students$$

3.3. Data Collection Procedure

A total of 255 questionnaires was administered to the respondents consisting of undergraduate students of JKUAT University (Eldoret Campus). Each respondent was required to complete the questionnaires in half an hour. The questions that were evaluated using questionnaires involved the analysis of various aspects in relation to classification of contraceptive use. The questionnaires were structured based on some closed-ended questions and a five-point Likert scaled data which determined various aspects on the use of contraceptives among undergraduate students.

3.4. Data Analysis

3.4.1. Chi-Squared Feature Selection

Feature selection process was performed based on Chi-Squared algorithm to obtain more relevant rankings based on significant features. The features are sorted in ascending order based on the Chi-Squared statistic values and used internally to perform classification. Features with higher Chi-Square value show that they are more independent on the response and it can be selected for model training. On the other hand, features with lower Chi-Square values show that they are highly dependent on the response and can be eliminated from the model training. Hence, the best and relevant are selected and used to perform classification while the rest of the features are eliminated from classification. The chi-square statistic is calculated as follows:

$$X^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{3}$$

Where $i = 1,2 \ldots \ldots \ldots n$

Oi and Ei are the observed and expected frequency respectively if these two categories are independent. If they are independent, then these O and E values will be close and if they have some association, then the Chi-squared value will be high. The x^2 statistic asymptotically follows the x^2 distribution with the degrees of freedom $(r - 1)\,(c - 1)$, where r is the number of categories for the independent variable and c is the number of categories for the dependent variable.

3.4.2. Classification of Contraceptive Use Using Decision Tree Classifier

(i). CHAID Algorithm

The CHAID (Chi-square automatic induction) method is employed in this study to categorize undergraduate students' use of contraceptives. CHAID differs from previous algorithms in that it uses the prepruning strategy, which aims

to halt branching before overfitting takes place. Another distinction is that because CHAID only uses categorical variables, ranges in continuous variables must be broken or classes must be substituted Rastogi, R., & Shim, K. [15]

The X^2 test is used by CHAID to determine whether to merge fields that do not result in statistically significant variations in the values of a target field. The insignificant categories are merged into a homogeneous group, and the remaining categories are analyzed repeatedly until the differences are no longer significant, as shown in the equation below. A Bonferroni adjusted p-value is computed for the merged cross tabulation as a result of these merging procedures. Then, if any of these splits provide a statistically significant difference in results, CHAID keeps that split. This process is repeated for every group having three or more fields Rastogi, R., & Shim, K. [15]. The field that produces the groups that differentiate the most, as determined by the X^2 test, is then selected as a splitter for that node. The tree continues to expand until there are no longer any splits that result in statistically significant categorization differences. The size of a tree and its usefulness as a classifier are determined by the precise level of significance.

$$x^2 = \sum_{i=1}^{l}\sum_{j=1}^{k}\frac{(x_{ij}-E_{ij})^2}{E_{ij}} \quad (4)$$

where Eij is the expected value.

(ii). C5.0 Algorithm

The C5.0 algorithm builds a decision tree model by calculating the best splits using information gain ratio (IGR) %. Pandya, R., & Pandya, J. [13]. The IGR involves a probability measure that is used to determine the level of uncertainty reduction. The decision tree is grown by calculating the split with the biggest IGR until the best solution is obtained. The IGR is calculated as shown in the equation below:

$$GainRatio = \frac{Gains(N,T)}{Ent(T)} \quad (5)$$

where Gain Ratio represents the IGR, N represents the global dataset, T represent the predictor variable, $Ent(T)$ represents entropy and $Gains(N,T)$ represents the entropy difference between the original and new nodes, and is calculated as follows:

$$Gains(N,T) = [\sum_i^t P(C_i|N)log_2P(C_i|N)] * \left(\sum_j^K \frac{|T_j|}{|N|} - 1\right) \quad (6)$$

where C is a set of target variable, t is the category number of C, K is the category number of T, $Ci\ (i = 1,2,\ldots,t), Tj\ (j = 1,2,\ldots,k)$.

The C5.0 algorithm involves the following basic steps. First, the process begins with initialization of the weights of the training sample. Then training subsets are obtained sequentially and the error of the subsets are calculated while updating weights. Finally, the training is ended and the classification results evaluated. Additionally, the verification accuracy of the model is evaluated using a cross-validation method. The dataset is divided into n folds, and one-fold is used for validation, and the other folds (n1 folds) are used as the training data Yao, *et al*. [28] The selection of every fold in the validation set results in n model accuracies being obtained through iteration of the same step. Finally, the average of the obtained accuracies known as the cross-validation accuracy (CAV) is considered the model accuracy. Therefore, this method is helpful for overfitting problems and improving the generalization capability of the decision tree model.

3.4.3. Performance of the Decision Tree Model

To determine the performance of the fitted model, the classification accuracy of the model is estimated using Confusion matrix. Confusion matrix gives detailed view of the performance with breakdown of correct and incorrect predictions for each class. The performance is measured by comparing the predicted outcome values with actual values as follows:

Table 1. *Confusion Matrix.*

Actual			
		Positive	**Negative**
Predicted Class	Positive	True positive count (TP)	False Positive count (FP)
	Negative	False Negative count (FN)	True Negative Count (TN)

This is useful in measuring the accuracy of the fitted model, Precision, Recall and Area-under-Curve (AUC-ROC). Overall accuracy is the rate at which the model makes accurate predictions. It is the ratio of number of correct predictions to total number of predictions made as shown in the equation below:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + TrueNegative + False\ Positive + False\ Negative} \quad (7)$$

In addition to accuracy, sensitivity, specificity, positive predictive value and negative predictive value were also estimated. Sensitivity (Recall) is used to measure how the model is used to identify events in the positive class as follows:

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (8)$$

Specificity is used to accurately identify negative ratios and can be presented in the form of a false positive rate as follows:

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive} \quad (9)$$

Positive predictive value is used to determine how a positive test delineates a true positive value as follows:

$$Positive\ Predictive\ Value = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (10)$$

Negative predictive value is used to determine the proportion of negative findings that can be associated with

true negative event and false negative event. It is calculated as follows:

$$Negative\ Predictive\ Value = \frac{True\ Negative}{True\ Negative + False\ Negative} \quad (11)$$

3.4.4. Goodness-of-Fit of the Decision Tree Model

Cohen's Kappa test was used to determine the goodness of fit of the fitted model. Kappa coefficient evaluates the accuracy of the models by measuring agreement between predicted values and true values. The range of Cohen's Kappa (k) value is limited to $\kappa \leq 1$ <=1. A Cohen's Kappa value<0 indicates no agreement, 0-0.20 indicates slight, 0.21-0.40 indicates fair, 0.41-0.60 indicates moderate, 0.61-0.80 indicates substantial and 0.81-1.00 indicates almost perfect agreement (Vieira et al., 2010). Cohen Kappa statistics is calculated as follows:

$$k = \frac{2*(TP*TN - FN*FP)}{[(TP+FP)*(FP+TN)]+[(TP+FN)*(FN+TN)]} \quad (12)$$

In addition, McNemar test was also used to evaluate the goodness of fit of the fitted model by determining whether the sample was drawn from the population that has equal proportions. The level of statistical significance was evaluated at 95% significance level. A statistically significant value implies that the sample population statistically differ across the dichotomous groups of the dependent variable. Hence the McNemar test is based on the following hypothesis:

H0: There is no statistical difference in predictive accuracy across the categories of the dichotomous variable.

H1: There is statistical difference in predictive accuracy across the categories of the dichotomous variable.

The McNemar test statistic is calculated as follows:

$$x^2 = \frac{(b-c)^2}{b+c} \quad (13)$$

4. Chapter 4: Data Analysis and Presentation

4.1. Feature Selection

Chi-squared feature selection was used to select the best and the most relevant features influencing contraceptives use among undergraduate students by ranking features in ascending order based on the values of Chi-squared scores. The first 15 features out of 34 were selected based on their Chi-Square scores while the rest were eliminated. The findings as shown in (*Table 2*) below present features that were ranked position 1 to 15 and their Chi-squared values.

Table 2. *Chi-Square feature Selection.*

Biomarker	Chi Square	Number Feature	Rank
Are you comfortable to purchase contraceptives in an open market	0.27681092	15	1
I use contraceptives because it is easily available in my place of residence	0.22463064	17	2
I use contraceptives because of religious influence	0.22029162	22	3
Have you ever had sexual impulses	0.21177213	13	4
I use contraceptives because I see friends using them	0.19454002	24	5
What is your general believe on use of contraceptives	0.19358137	14	6
I don't use contraception because I did not prepare the pills for unplanned sex	0.17847613	30	7
I use contraceptives because of fear of getting pregnant	0.16546599	23	8
I use contraceptives because they are dispensed in the institution	0.16226783	20	9
I use contraceptives because I have multiple partners	0.15786698	21	10
Which department are you in	0.15283481	5	11
I don't use contraception because my partner do not want me using them	0.13637125	33	12
I use contraceptives because I am not employed	0.13604251	28	13
Which contraceptives method do you think is suitable for college students	0.12693567	8	14
Which diseases does contraceptives prevent	0.12322895	7	15

4.2. Decision Tree Classifier

Decision tree classifier was fitted to classify contraceptives use among undergraduate students. Two models were built for classification where the first model was build based on CHAID algorithm while the second model was build based on C5.0 algorithm. The models include one dependent variable which is the frequency of contraceptives involving a multi-label classification with three classes including always, sometimes and occasional.

4.2.1. CHAID Algorithm

The decision tree algorithm based on CHAID algorithm was conducted to produce a decision tree output which is used to predict overall probability of contraceptives use among undergraduate students. The findings show that the structural model constructed using the CHAID algorithm exhibited three layers of attributes (*Figure 1*). As observed, the CHAID algorithm classifies five factors that are considered significant for influencing contraceptives use. These included the ease of availability in the place of residence, the general believe about contraceptives by students, sexual impulses, having multiple partners and the fact that they are not employed. These factors significantly influence the frequency of contraceptives use among undergraduate students.

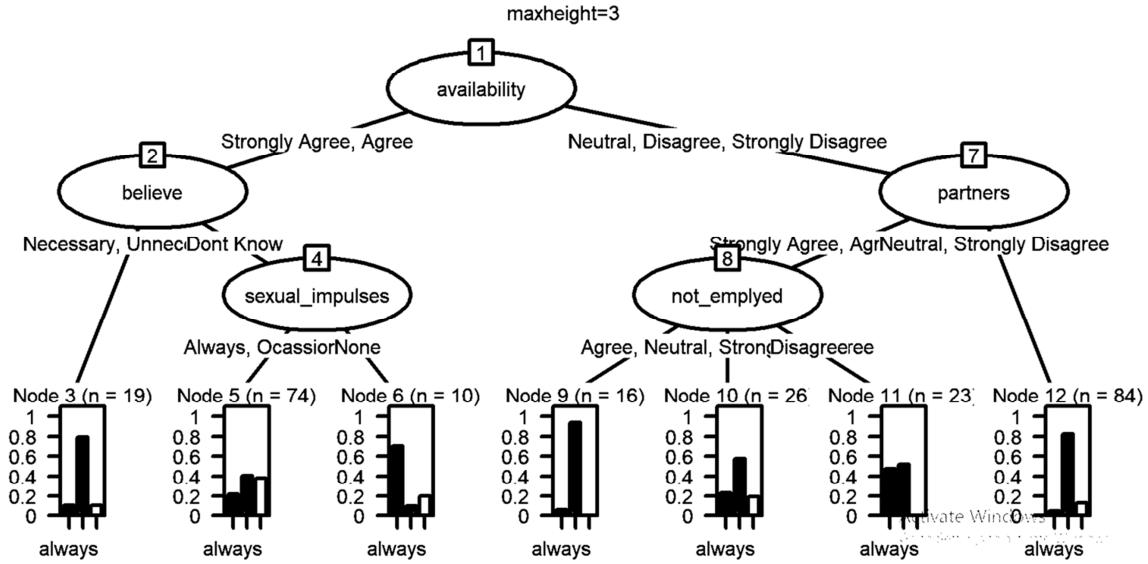

Figure 1. *Structural model of the CHAID algorithm (Source; Author).*

4.2.2. C5.0 Algorithm

The decision tree model constructed using C5.0 algorithm exhibited five layers of attributes. The findings as shown in (*Figure 2)* below show the C5.0 decision tree until node 13. From C5.0 algorithm, it is clear that the most significant factors that were classified to influence contraceptives use among undergraduate students included having sexual impulses, ease of availability in their place of residence, fear of getting pregnant, unplanned sex and the general believe on contraceptives use.

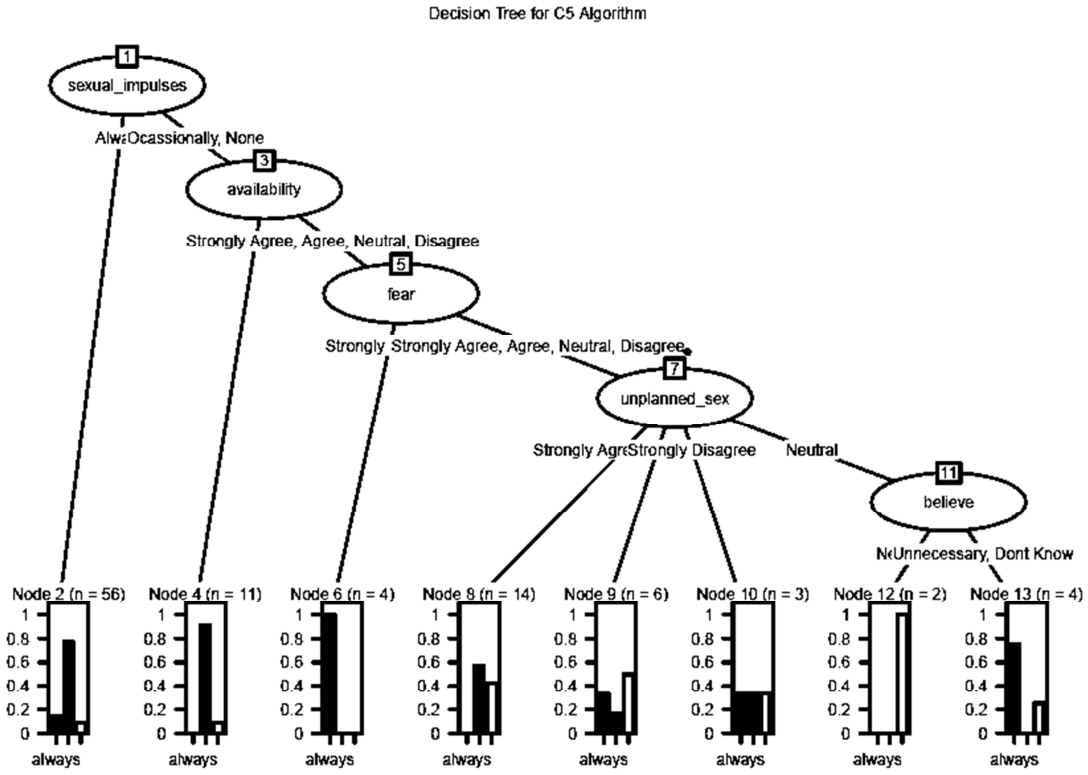

Figure 2. *Decision Tree model with C5.0 Algorithm (Source; Author).*

4.2.3. Performance of the Decision Tree Model

The predictive performance of the decision tree algorithm was evaluated based on performance measures such as confusion matrix, overall accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value

(NPV). The data set has a total of 252 observations. Confusion Matrix of the decision tree classifiers are shown in Table 3. and Table 4. below, where the findings show that the total number of correct predictions are observed to be 195 (26+145+24) and 163 (7+156+0) for the CHAID Algorithm & the C5.0 Algorithm respectively. The majority class is represented by the second class involving those who sometimes use contraceptives.

Table 3. *Confusion Matrix of the CHAID Algorithm.*

		Actual			**Total**
		Always	**Sometimes**	**Occasional**	
Predicted	Always	26	4	3	33
	Sometimes	15	145	21	181
	Occasional	6	8	24	38
Total		47	157	48	252

Table 4. *Confusion Matrix of the C5.0 Algorithm.*

		Actual			**Total**
		Always	**Sometimes**	**Occasional**	
Predicted	Always	7	1	2	10
	Sometimes	40	156	46	242
	Occasional	0	0	0	0
Total		47	157	48	252

The findings as shown in Table 5. below suggest that the decision tree classifier based on the first model (CHAID Algorithm) resulted in an overall data accuracy of 64.68% with 95% confidence interval of between 58.55% and 70.58% in predicting contraceptives use. On the other hand, the second model (C5.0 Algorithm) showed an accuracy of 61.18% in predicting contraceptives use with 95% CI between 52.95% and 68.97% (Table 5.).

Table 5. *Performance accuracy of the fitted Decision tree models.*

	Accuracy	**Sensitivity**	**Specificity**	**Positive Predictive Value**	**Negative Predictive Value**
	First model based on CHAID Algorithm				
Overall Accuracy (95% CI)	0.6468 (0.5844, 0.7058)				
Kappa's coefficient	0.1043				
McNemar	<2e-16				
First class (always)	0.56715	0.14894	0.98537	0.7000	0.83471
Second class (sometimes)	0.54418	0.99363	0.09474	0.64463	0.9000
Third class (occasional)	0.50000	0.0000	1.0000	-	0.8095
	Second Model based on C5.0 Algorithm				
Overall Accuracy (95% CI)	0.6118 (0.5295, 0.6897)				
Kappa's coefficient	0.1464				
McNemar Test	5.747e-05				
First class (always)	0.55873	0.20690	0.91057	0.35294	0.82963
Second class (sometimes)	0.5847	0.8936	0.2759	0.6667	0.6154
Third class (occasional)	0.52733	0.10345	0.95122	0.33333	0.81818

Therefore, the findings suggest that the decision tree classifier based on the first model build with CHAID algorithm depicted a higher level of performance as compared to the second one with C5.0 algorithm.

4.2.4. Goodness of Fit of the Decision Tree Classifier

The findings as shown in (Table 5.) above suggest that the Kappa's coefficient for the first model was 0.1043 while for the second model was 0.1464. As observed, the Kappa's values for both the first and second models are between 0.0-0.20 implying a slight agreement. Therefore, these empirical findings depict good consistency of the model between the predicted generated and the actual true values. In addition, the findings based on McNemar test suggest a p-value of <2e-16 in model 1 and 5.747e-05 in model 2 (Table 5.) These p-values are less than 0.05 implying a statistically significant result hence the proportion of contraceptives among undergraduate students statistically differ across the three categories of the dichotomous dependent variable. Thus, this indicate that the use of contraceptives among undergraduate students is influenced by significant factors which suggest that the fitted decision tree model is good and reliable.

5. Chapter 5: Summary, Conclusion and Recommendation

5.1. Summary of the Findings

The aim of the study was to classify contraceptives use among undergraduate students based on a sample of undergraduate students at JKUAT University (Eldoret Campus). A decision tree classifier based on CHAID & C5.0 algorithm was used to achieve the research objectives.

Feature selection based on Chi-Squared algorithm was used to rank important factors influencing contraceptives use among undergraduate students. These features were then applied to fit the decision tree classifier. The findings show that factors influencing contraceptives use that were ranked higher by Chi-Squared feature selection method had higher classification performance in the fitted decision tree model.

The performance of the model was evaluated based on classification accuracy, confusion matrix, sensitivity, specificity, positive predictive value and negative predictive value. The findings show that the CHAID decision tree model achieved a higher accuracy of 64.68% as compared to the C5.0 decision tree model with accuracy of 61.18%. Many studies have compared the performance of the decision tree based on different decision tree algorithms. Priyam, *et al*. [31] Predicted performance of students suing both SPRINT and SLIQ decision tree classifier where SPRINT algorithm was found to have a higher accuracy as compared to SLIQ algorithm. The findings achieved an accuracy of 98.68% and 98.72% respectively Azar, *et al*. [27] used three types of decision tree classifiers to detect breast cancer based on single decision tree, boosted decision tree and decision tree forest. The findings show that boosted decision tree achieved a higher accuracy of 98.83% with 437 correct classifications in the training phase as compared to single decision tree with accuracy of 97.07% with 429 correct classifications. However, during validation phase, decision tree forest achieved a higher accuracy of 97.51% as compared to both boosted and single decision trees classifiers with accuracy of 97.07% and 95.75% respectively Ibad, *et al*. [10] achieved an accuracy of 88.99% in the classification of contraceptives use in BKKBN based on Information Index decision tree classifier. Lee, S., *et al*. [24] Compared the performance of Rao-Stirling index and Gini index in the classification of distances between classes using decision tree classifier. The study found out that Rao-Stirling index achieved a higher accuracy as compared to Gini index.

However, the C5.0 decision tree model achieved a higher Kappa's value of 0.1464 as compared to the CHAID decision tree with Kappa value of 0.1043. Nevertheless, these two values suggest a slight agreement implying good consistency of the predicted values and true values. Furthermore, McNemar test depicted a statistically significant associations of the selected factors and the dependent variables. This implies that the proportions of contraceptives use statistically differ across the three categories of the dependent variable hence the fitted decision tree classifier depict a good fit for predicting contraceptives use among undergraduate students. This aligns the findings of Tangirala, S. [18] who found that the goodness of the split based on GINI index and Information gain individually achieved the same accuracy of 78.5%.

5.2. Recommendations

Since this study was carried at out at JKUAT, there was limitation of time and data. Hence, future study should be carried out with data set representing larger population so as to obtain more viable information on contraceptive use. It is recommended that students should be made aware on contraceptive use and family planning methods. Also, other Machine learning Techniques can be explored.

References

[1] Ahmed, S., Li, Q., Liu, L., & Tsui, A. O. (2012). Maternal deaths averted by contraceptive use: an analysis of 172 countries. *The Lancet*, *380* (9837), 111-125.

[2] Brunner Huber, L. R., Smith, K., Sha, W., & Vick, T. (2018). Interbirth interval and pregnancy complications and outcomes: findings from the pregnancy risk assessment monitoring system. *Journal of Midwifery & Women's Health*, *63* (4), 436-445.

[3] Chiang, H. J., Tseng, C. C., & Torng, C. C. (2013). A retrospective analysis of prognostic indicators in dental implant therapy using the C5. 0 decision tree algorithm. Journal of Dental Sciences, 8 (3), 248-255.

[4] Collis, J. & Hussey, R. (2013) Business Resarch. *England: Palgrave Macmillan*.

[5] Darroch, J. E., & Singh, S. (2011). Estimating unintended pregnancies averted from couple-years of protection (CYP). *New York: Guttmacher Institute*, *1* (10).

[6] Darroch, J. E., Woog, V., Bankole, A., & Ashford, L. S. (2016). Adding it up: costs and benefits of meeting the contraceptive needs of adolescents.

[7] Dauda, K. A., Babatunde, A. N., Olorede, K. O., Abdulsalam, S. O., & Ogundokun, O. R. (2018). Effectiveness of Contraceptive Usage among Reproductive Ages in Nigeria Using Artificial Neural Network (ANN). *Computing & Information Systems*, *22* (1).

[8] Ftye, M., Letta, A., & Achamyeleh, B. (2022). designing a predictive model for the likelihood of contraceptive method usage in ethiopia. *Cosmos Journal of Engineering & Technology*, *12* (1).

[9] Haq, I., Hossain, M. I., Rahman, M. M., Methun, M. I. H., Talukder, A., Habib, M. J., & Hossain, M. S. (2022). Machine Learning Algorithm-Based Contraceptive Practice among Ever-Married Women in Bangladesh: A Hierarchical Machine Learning Classification Approach. In *Machine Learning and Data Mining-Annual Volume 2022*. IntechOpen.

[10] Ibad, M., Lutfiya, I., Handayani, D., Fasya, A. H. Z., Herowati, D., & Sari, M. P. (2023, May). Classification analysis using decision tree on factors that influence the selection of contraception equipment in East Java Province. In *AIP Conference Proceedings* (Vol. 2595, No. 1). AIP Publishing.

[11] Jiang, W., & Ha, L. (2020). Smartphones or computers for online sex education? A contraception information seeking model for Chinese college students. *Sex Education*, *20* (4), 457-476.

[12] Mustaqim, B. W., & Surarso, B. (2020). combination of synthetic minority oversampling technique (smote) and backpropagation neural network to contraceptive iud prediction.

[13] Pandya, R., & Pandya, J. (2015). C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning. International Journal of Computer Applications, 117 (16), 18-21.

[14] Podolskyi, V., Gemzell-Danielsson, K., & Marions, L. (2018). Contraceptive experience and perception, a survey among Ukrainian women. *BMC Women's Health*, *18*, 1-6.

[15] Rastogi, R., & Shim, K. (2000). PUBLIC: A decision tree classifier that integrates building and pruning. *Data Mining and Knowledge Discovery*, *4*, 315-344.

[16] Saunders, M. N., Thornhill, A. & Lewis, P. (2019). *Research Methods for Business Students* (Eighth Edition ed.). London, UK: Pearson.

[17] Sun, X., Liu, X., Shi, Y., Wang, Y., Wang, P., & Chang, C. (2013). Determinants of risky sexual behavior and condom use among college students in China. *AIDS care*, *25* (6), 775-783.

[18] Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, *11* (2), 612-619.

[19] UNFP (2010). United Nations Population Fund. Sexual and Reproductive Health for all: Reducing Poverty, Advancing Development and Protecting Human Rights. *New York, New York*, *United States*.

[20] Vieira, S. M., Kaymak, U., & Sousa, J. M. (2010, July). Cohen's kappa coefficient as a performance measure for feature selection. In International conference on fuzzy systems (pp. 1-8). IEEE.

[21] Wang, H., Long, L., Cai, H., Wu, Y., Xu, J., Shu, C.,... & Yin, P. (2015). Contraception and unintended pregnancy among unmarried female university students: a cross-sectional study from China. *PloS one*, *10* (6), e0130212.

[22] Wang, Y., Chen, M., Tan, S., Qu, X., Wang, H., Liang, X. & Tang, K. (2020). The socioeconomic and lifestyle determinants of contraceptive use among Chinese college students: a cross-sectional study. *Reproductive Health*, *17*, 1-11.

[23] Zhou, H., Wang, X. Y., Ye, F., Gu, H. H., Zeng, X. P. L., & Wang, Y. (2012). Contraceptive knowledge, attitudes and behavior about sexuality among college students in Beijing, China. *Chinese medical journal*, *125* (06), 1153-1157.

[24] Lee, S., Lee, C., Mun, K. G., & Kim, D. (2022). Decision Tree Algorithm Considering Distances Between Classes. *IEEE Access*, *10*, 69750-69756.

[25] Jin, L., & Myers, S. C. (2006). R2 around the World: NNew Theory and New Tests. Journal of Financial Economics, 79, 257-292.

[26] Kuang.. &Davison (2017). Learning Word Embeddings with Chi-Square Weights for Healthcare Tweet Classification. -2017/08/17.10.3390/app7080846.

[27] Azar, A. T., & El-Metwally, S. M. (2013). Decision tree classifiers for automated medical diagnosis. *Neural Computing and Applications*, *23*, 2387-2403.

[28] Yao, J. G., Weasner, B. M., Wang, L. H., Jang, C. C., Weasner, B., Tang, C. Y., Salzer, C. L., Chen, C. H., Hay, B., Sun, Y. H., Kumar, J. P. (2008). Differential requirements for the Pax6 (5a) genes eyegone and twin of eyegone during eye development in Drosophila. Dev. Biol. 315 (2): 535--551.

[29] A, Rehman., Abbas, A. Chandio., I. Hussain., L, Jingdong, (2019). Fertilizer consumption, water availability and credit distribution: Major factors affecting agricultural productivity in Pakistan, *Journal of the Saudi Society of Agricultural Sciences*, Volume 18, Issue 3, 2019, Pages 269-274.

[30] Njoroge, p. (2016). Factors Influencing Uptake of Contraceptive Services Among Undergraduate Students aged 18-35 years at jomo kenyatta university of agriculture and technology, kenya. *A Master Thesis in Public Health in the Jomo Kenyatta University Of Agriculture and Technology.*

[31] Priyam, A., Abhijeeta, G. R., Rathee, A., & Srivastava, S. (2013). Comparative analysis of decision tree classification algorithms. *International Journal of current engineering and technology*, *3* (2), 334-337.

[32] Kithuka, B (2012). Factors Associated with Condom Use among Students at Jomo Kenyatta University of Agriculture and Technology. *A Master of science in Epidemiology in the Jomo Kenyatta University of Agriculture & Technology.*