© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

Using data mining techniques to determine variables influencing condom usage versus non-usage in a sample of sexually active adolescents in South Africa

R. J. Blignaut[1], T. Vergnani[2]
[1]*Dept of Statistics,* [2]*Faculty of Education,*
University of the Western Cape, South Africa.

Abstract

This study aims to identify the variables that influence condom usage among a random sample of grade 11 secondary school pupils in the Cape Town metropolitan area. In order to find the variables that most influenced condom usage on the last coital episode, data mining techniques were applied using the SAS Enterprise Miner program. Various decision tree algorithms were implemented and the results with various program option settings showed that a variable called 'construct availability' (namely either talking about condoms with a sexual partner or thinking about condoms) most consistently distinguished between condom users and non-users. Other variables shown to distinguish between condom users and non-users were intention to use condoms, self-efficacy, self-standards, condom availability, attitudes, AIDS worry, norms, race and gender. These variables need to inform both the marketing of condoms to young people as well as life-skills programmes in schools.

1 Introduction

The major public health problem facing South Africa in the next decade is to curb the rapid spread of HIV infection. According to a recent UNAIDS report [6], 19.9% of adults in South Africa are now infected with HIV, which is an increase from 12.9% two years ago. "With a total of 4.2 million infected people, South Africa has the largest number of people living with HIV/AIDS in the world" (UNAIDS [6]). According to Whiteside and Sunter [7], adult prevalence

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

rates could be as high as 30% by the year 2010, which would imply a doubling of the present number of those infected with HIV. The epidemic in South Africa is characterised by mainly heterosexual transmission and by an extremely rapid spread amongst adolescents and young adults in their early twenties.

It is generally accepted, that given the lack of vaccine against HIV, one of the best means of protection against the virus is the correct and consistent use of condoms. However, research shows that adolescent populations worldwide as well as in South Africa show low, inconsistent and/or infrequent use of condoms (Kelly, [2]; Lovelife &Henry J. Kaiser Foundation, [3]).

This study aims to identify the key psycho-social variables that influence condom usage amongst a random sample of grade 11 secondary school pupils in the Cape Town metropolitan area.

2 Methods

Data were collected by means of self-administered questionnaires. The sample comprised a random sample of 1931 grade 11 learners from the Cape Town metropolitan area. Only 1884 questionnaires could be used in the analyses after tests for consistency in answers were done. The three most prevalent racial groups living in the area were included in the sample namely: Black, White and Coloured (of mixed racial decent). The independent variables were derived from four of the major psycho-social theories used to explain behaviour, namely, the Health Belief Model, Social Learning Theory, Theories of Reasoned Action and Planned Behaviour, Theory of Subjective Culture and Interpersonal Relations, as well as the construct availability model. The independent variables were intention to use condoms, beliefs, attitudes, norms, self-standards, affect, construct availability, condom availability, AIDS worry, perceived vulnerability to AIDS, and perceived self-efficacy.

In order to find the variables that best influenced condom usage on the last coital episode, data mining techniques were applied using the SAS Enterprise Miner program [5]. Stepwise variable selection and logistic regression procedures as well as various decision tree algorithms were derived to determine the variables that best differentiated between condom users and non-users.

The focus of this study is on the application of decision tree algorithms. An advantage of decision tree algorithms over other modelling procedures is that a model is derived in interpretable English rules or logic statements. Another advantage is that missing values can also be used as input when searching for splitting rules in decision tree algorithms. When a tree is constructed the original node (grouping) with all data is referred to as the root node. The node with all successors forms a branch and the final node is referred to as a leaf.

In the following section various decision tree algorithms, namely, the SAS tree procedure, CHAID (Chi-squared Automatic Interaction Detection), CART (Classification and Regression Trees), C4.5 and C5.0 will briefly be discussed (Potts [4]). An analysis of the similarities and differences in these techniques will provide some understanding as to why results may differ when applying

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

these techniques to this data set. The application of these algorithms will be discussed in the results section of this article.

The SAS implementation of decision trees finds multi-way splits based on nominal, ordinal, and interval inputs. The splitting criteria and options that determine the method of tree construction can be selected in the program. The criterion for evaluating a splitting rule may be based on either a statistical significance test, namely an F test or a Chi-square test, or on the reduction in variance, Entropy, or the Gini impurity measure. The F test and Chi-square test accept a p-value input as a stopping rule. When the Chi-square splitting criteria is used the statistic computed is the Log-worth which is equal to –log(p-value from the Chi-square). For the Gini or Entropy criteria the computed statistic is the Worth, which measures reduction in variance for the split. The Worth is calculated by summing over b branches, of P(b) times I(b), where I(b) denotes the Entropy, Gini, or variance measure in the node, and P(b) denotes the proportion of observations in the node assigned to branch b. For a specific node the SAS tree node seeks the split with maximum Worth or Log-worth subject to the limit on the number of branches and the limit on the minimum number of observations assigned to a branch. By default, p-values are adjusted to take into account multiple testing. The Kass adjustment before the split results in selecting the best split where the smallest Kass adjusted p-value occurs. The SAS tree can be grown automatically or interactively.

The SAS tree program can approximate decision tree algorithms such as CHAID (Chi-squared automatic interaction detection), CART (Classification and Regression Trees) and C4.5.

The CHAID method of tree construction stops when a specified Chi-square significance level is reached. CHAID attempts to stop growing the tree before over-fitting occurs. The inputs are either nominal or ordinal. The CHAID exhaustive method is similar to the SAS tree node's heuristic method. The CHAID algorithm differs from the SAS tree algorithm in a number of ways:

- The SAS tree node seeks the split minimizing the adjusted p-value.
- The CHAID algorithm cannot use interval variables.
- The SAS tree node cannot approximate the CHAID method for an ordinal target.

For the C4.5, C5.0 and CART methods it is argued that the right thresholds for stopping the tree construction are not known in advance and therefore over-fitting is recommended followed by the pruning of some less significant nodes.

The CART algorithm constructs a binary tree and evaluation is done in terms of an overall Gini index for nominal targets and variance reduction for interval targets.

The main difference between C4.5 and CART is that CART performs a binary split at each node and produces a binary tree whereas C4.5 produces a tree with any number of branches depending on the number of variable categories. For C4.5, the target is nominal and the inputs may be nominal or interval. The recommended splitting criterion is the reduction in Entropy. For interval inputs, C4.5 finds the best binary split. For nominal inputs, a branch is created for every category, and then, optionally, the branches are merged until the splitting

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

measure does not improve. The tree is grown to over-fit the training data. C4.5 and CART also differ with respect to the method of pruning of the tree (Berry [1]).

The C5.0 algorithm is an improvement of the C4.5 algorithm with misclassification costs and cross-validation included. The C5.0 and the SAS tree node differ mainly because C5.0 creates binary splits on interval inputs and multi-way splits on nominal inputs, whereas the SAS tree node treats interval and nominal inputs the same. For the SAS tree node a validation data set needs to be present to implement pruning whereas C5.0 does not have such restriction.

3 Results

3.1 Descriptive information

In total 1016 (53.9%) of the learners sampled were sexually active. Of the sexually active group 46 percent were male and 54 percent were female. The average age of this group was 17.86 years (standard deviation=1.746, minimum=14 and maximum=30). Forty-five percent of this group used a condom during their last coital episode and sixty-seven percent had condoms available.

In the next section the results of various decision tree algorithms will be shown. The algorithms will aim to grow trees that best distinguish between condom users and non-users on the last coital episode.

3.2 Decision tree analyses

The architecture of the various SAS tree algorithms applied is summarized in Table 1. The SAS program specifies program settings to approximate CHAID, CART and C4.5 algorithms. For all trees the minimum number of observations per leaf was set to 10 and the maximum number of observations required for a split search was 20. Trees were all grown to a depth of 6. The original data were divided into a training data set (70%) and a validation data set (30%). The training data set is used to train the algorithm and the validation data set is used to validate the tree rules for classification. The model assessment finds the best tree based on the results obtained from the validation data set. Either the average square error (ASE) or the total leaf impurity (Gini index) (TLI) was used. Kass p-value adjustments were done before the number of branches were chosen for the SAS Chi-square tree and the C4.5 algorithms. For these algorithms p-value adjustments were also done for the depth of the tree. The CHAID algorithm used the Kass p-value adjustment after choosing the number of branches.

3.2.1 SAS Chi-square tree algorithm

When using the SAS Chi-square test as the splitting criteria the tree grown only had 3 leaves. When testing the tree, the percentage of observations correctly classified, when using the validation data set, was 66.9 percent. From the tree diagram (Figure 1) it can be seen that learners who were likely to use condoms

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

had high "Construct availability" scores (between 7 and 10) as well as a high "Intention to use condoms" score of 10. The variables competing for the first split were "Construct availability" (Log-worth=15.496) followed by "Self-efficacy" (Log-worth=11.697), "Self-standards" (log-worth=11.643), "Intention to use condoms" (Log-worth=10.763) and "Condom availability" (Log-worth=7.021).

Table 1: SAS tree algorithms applied.

	SAS Chi-square	*SAS Gini reduction*	*SAS Entropy reduction*	*CHAID approx.*	*CART approx.*	*C4.5 approx.*
Splitting criteria	Chi-square	Gini	Entropy reduction	Chi-square	Gini	Entropy reduction
p-value	0.2	-	-	0.05	-	-
Max nr of branches	2	2	2	2	2	10
Surrogate rules saved	0	0	0	0	5	0
Missing values used	No	No	No	Yes	No	No
Model assessment	ASE	ASE	ASE	TLI	TLI	ASE
Correctly classified (%)	66.9	67.2	69.2	37.3	34.7	69.8
Final Leaves	3	5	10	21	38	23

When inspecting the Chi-square tree (see Figure 1) only 46 percent (code=1, first node in the second column) of the learners in the training data set used condoms on the last coital episode. When only high "Construct availability" scores (7-10) were used the condom usage increased to 59 percent in the training data set (second level, second column) and 53.3 percent in the validation data set (second level, third column). When the additional constraint of a high "Intention to use condoms" score was added, the condom usage increased to 71.4 percent for the training data set (third level, second column) and 64.9 percent for the validation data set (third level, third column). With "Construct availability" scores less than 7, only 27 percent of the learners used condoms (second level, first leaf).

3.2.2 SAS tree with Gini reduction as splitting criteria

When the Gini reduction was used instead of the Chi-square test as the splitting criteria a 5 leaf tree was trained with a 67.2 percent correct classification on the validation data set. This tree differed with respect to the competing splits for the first node compared to all the other tree algorithms. "Self-efficacy" (Worth=0.116) was the most important variable selected for the first split.

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

Variables competing for the first split were "Attitudes" (Worth=0.092), "Intention to use condoms" (Worth=0.038), "Aids worry" (Worth=0.037) and Race (Worth=0.026). This tree (see Figure 2) showed that if the "Self-efficacy" scores were higher than 35.5 and "Intention to use condoms" had a score of 10, sixty-two percent of the learners would use condoms when testing the validation data on the trained tree. If the "Self-efficacy" score was increased to more than 43.5 the condom usage (in the validation data set) increased to 69.5 percent. When "Self-efficacy" scores between 35 and 43.5 were measured, the condom usage for learners with high "Construct availability" scores was 59.5 percent when validating the trained tree.

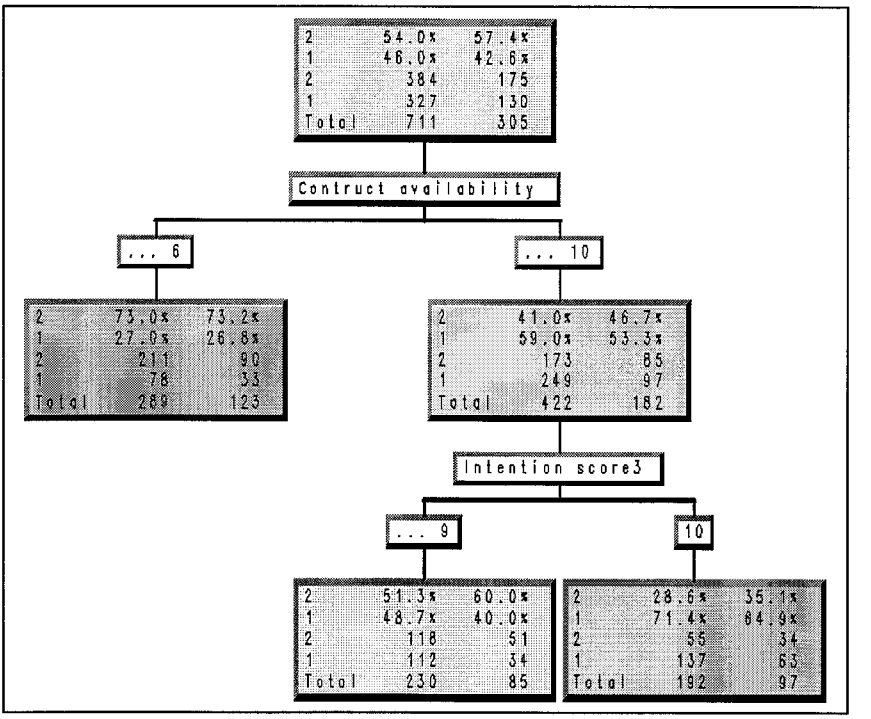


Figure 1: SAS Chi-square tree for predicting condom usage.

3.2.3 SAS tree with Entropy reduction as splitting criteria

With the introduction of Entropy reduction as the splitting criteria rule a tree was trained with 10 leaves and correct classification of the validation data of 69.2 percent. This tree created two branches, one with a reasonably good prediction for condom usage (see Figure 3). The first node was split on "Construct availability". The one branch developed where the "Construct availability" score was 7 or more. Furthermore if "Intention to use condoms" and the "Norms" scores were high, the condom usage was 64.3 percent when testing the validation data on the trained tree. The other branch showed (not shown on the tree

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

diagram) that if "Construct availability" was lower than 7 but "Attitudes" and "Self-efficacy" scores were high; condoms would be used if they were available.

3.2.4 CHAID approximation using SAS

The CHAID algorithm trained a tree with 21 leaves. On testing this tree it classified only 37.3 percent of the observations correctly. According to this tree "Construct availability" was selected as the most important variable (Log-worth=15.496) to predict the target. The other variables competing for the first split were in order of importance: "Self-efficacy" (Log-worth=11.735), "Self-standards" (Log-worth=11.545); "Intention to use condoms" (Log-worth=10.762), and then "Condom availability" (Log-worth=7.021).

According to this tree, learners would use condoms with high (9-10) "Construct availability" scores followed by high intention and normative scores as well as the availability of condoms. The classification using this branch improved further for the White and Coloured learners with high "Attitude" scores. From the racial node split it could be seen that Black female learners were more inclined to use condoms compared to their male counterparts.

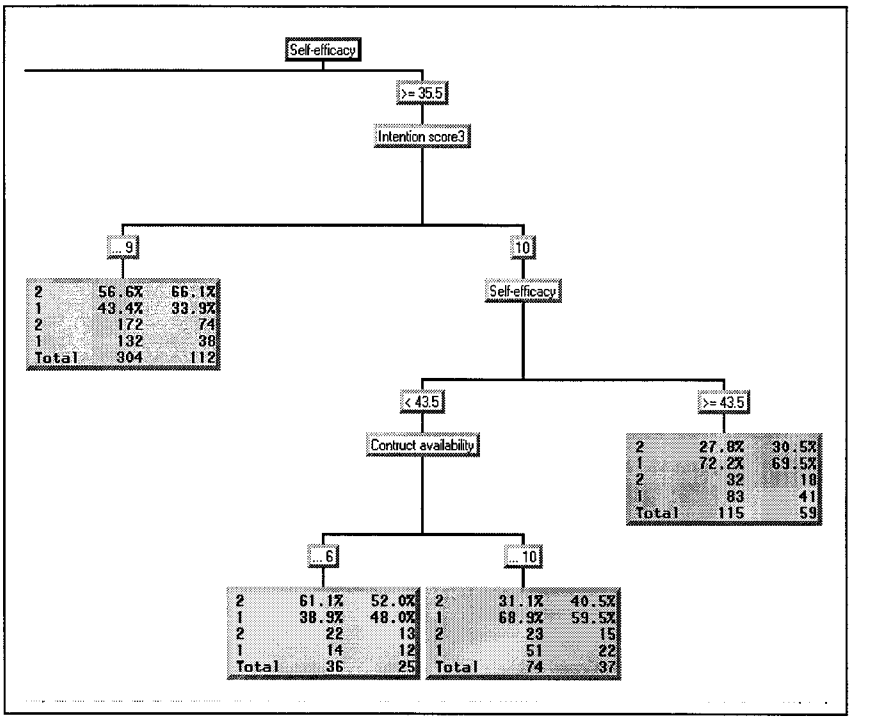


Figure 2: SAS Gini reduction tree for predicting condom usage.

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

3.2.5 CART approximation using SAS

The CART algorithm developed a tree with 38 leaves but the correct classification of the validation data set was only 34.7 percent. The initial split was on "Construct availability" (Worth=0.049). The competing variables for the first split were: "Self-efficacy" (Worth=0.04); "Self-standards" (Worth=0.037); "Intention to use condoms" (Worth=0.034) and "Attitudes" (Worth=0.032). This tree indicated that learners would use condoms when high "Construct availability", "Intention to use condom" and "Norms" scores were present. If condoms were available the Black learners reported a slightly lower use of condoms. For the White and Coloured learners the usage of condoms increased if the "Attitudes" scores were high.

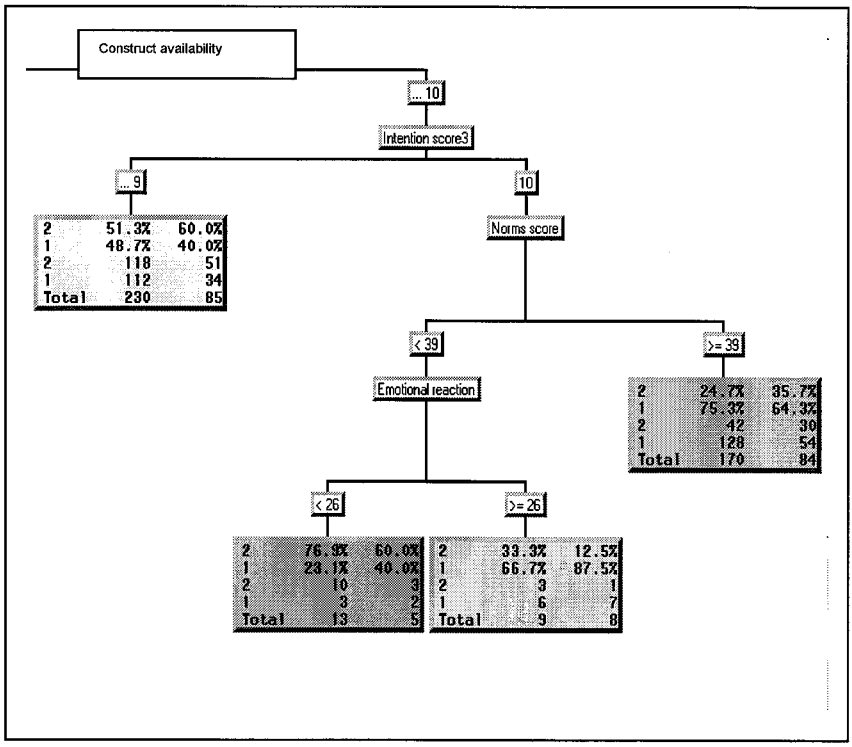


Figure 3: SAS Entropy reduction tree for predicting condom usage.

3.2.6 C4.5 approximation using SAS

The C4.5 algorithm trained a tree with 23 leaves and a correct classification of 69.8 percent when using the validation data. "Construct availability" (Worth=0.106) was again selected for the first split. Variables competing for the first split were: "Intention to use condoms" (Worth=0.082), "Self-efficacy" (Worth=0.076), "Self-standards" (Worth=0.067) and "Attitudes" (Worth=0.052).

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

A "Construct availability" score of 10 would result in condom usage of 69.5 percent. If "Construct availability" was 9 and "Self-efficacy" was high, condom usage would also be high (60%). If the 'Construct availability' score were below 6, condom usage would be most likely to take place if the general attitude and AIDS worry scores were high.

3.3 Logistic regression and variable selection results

As this article focuses on the use of decision tree algorithms, results of other modelling procedures will not be discussed in detail. It is interesting to note that the most imported variables selected by variable selection or stepwise logistic regression procedures were the same variables that competed for the first split in the decision trees.

4 Discussion and conclusion

It should be noted that the SAS program was used to develop all the tree algorithms. Specific program settings as specified by the SAS program were used to obtain approximations for CHAID, CART and C4.5.

When comparing the SAS Chi-square and CHAID algorithms (see Table 1), note that the architecture differed with respect to the *p*-value specified, the model assessment method and the fact that CHAID could use missing values as an acceptable value. (It should be noted that most of the variables included as input data to the tree algorithms had no missing values due to imputation prior to construction of the psychometric score variables). The SAS Gini reduction algorithm differed from the CART approximation with respect to the model assessment method used. The ASE model assessment approach produced better decision rules for classification for this problem compared to the TLI assessment. The SAS Entropy reduction method was similar to the C4.5 approximation except for an increase in branches allowed by the C4.5 algorithm. The increase in branches resulted in a tree with more leaves.

On inspection of the trees it can be seen that "Construct availability" was selected as the most important variable for splitting the first node in all the trees except for the SAS Gini reduction algorithm where "Self-efficacy" was selected. "Construct availability" and "Intention" were selected as variables that could influence the use of condoms in all the trees constructed. Other variables selected to favour condom usage were: "Self-standards", "Attitudes", "AIDS worry", "Norms", "Condom availability", race and gender. The order of importance of the variables selected varied from tree to tree.

The algorithms that fit a model with the highest proportion of correctly classified observations from the validation data set were the trees using Entropy reduction as the splitting criteria (SAS Entropy reduction and C4.5). The SAS Entropy reduction algorithm trained a simpler tree (10 leaves) and is to be favoured above the more complex C4.5 tree (23 leaves). The SAS Chi-square algorithm produced the simplest tree with reasonable accuracy for classifying observations from the validation data set.

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

As "Construct Availability" (talking about condoms with a partner or thinking about condoms) was selected as the most important variable to distinguish between condom users and non-users, life-skills curricula for schools should be designed to include the practicing of skills necessary to enable young people to talk about and negotiate condom usage with their sexual partners. In addition, life-skills programmes should aim at enhancing self-esteem and fostering positive attitudes towards condoms. Social marketing programmes should focus on fostering the use of condoms as part of a positive self-standard: for example, messages such as 'real men wear condoms' could help make condoms part of self-standards. The research also highlights the need to make condoms easily accessible to adolescents.

We hereby thank SAS South Africa and SAS Europe for the use of the SAS Enterprise Miner program for teaching and research activities. For financial support we thank the University of the Western Cape and the National Research Foundation (NRF). Thanks to the Western Cape Education department and the learners who participated in the study.

References

[1] Berry, M.J.A, Linoff, G. *Data Mining Techniques for Marketing, Sales and Customer Support,* Wiley Computer Publishers, New York, pp. 243-285, 1997.

[2] Kelly, K. *Communicating for Action: A contextual evaluation of youth responses to HIV/AIDS.* Sentinel Site Monitoring and Evaluation Project. Summary of Findings. Department of Health: Beyond Awareness Campaign, HIV/AIDS and STD Directorate, 2000.

[3] Lovelife & Henry J. Kaiser Family Foundation. *Hot Prospects, Cold Facts: Portrait of Young South Africa.* Parklands, Lovelife, 2001.

[4] Potts, W.J.E. *Decision Tree Modelling Course Notes.* SAS Institute Inc., Cary, NC 27513, USA, 2001.

[5] SAS Institute Inc. *SAS Enterprise Miner version 4.1.* Cary, New York, 1999-2000.

[6] UNAIDS. *Report on the global HIV/AIDS epidemic. June 2000.* Geneva, UNAIDS, 2000.

[7] Whiteside, A. & Sunter, C. *AIDS. The Challenge for South Africa.* Cape Town, Human & Rousseau, 2000.