PACIS 2010 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

2010

Smokers' Characteristics and Cluster Based Quitting Rule Discovery Model for Enhancement of Government's Tobacco Control Systems

Shamsul Huda
University of Ballarat, s.huda@ballarat.edu.au

John Yearwood
University of Ballarat, j.yearwood@ballarat.edu.au

Ron Borland
The Cancer Council Victoria, Ron.Borland@CancerVic.org.au

Recommended Citation

Huda, Shamsul; Yearwood, John; and Borland, Ron, "Smokers' Characteristics and Cluster Based Quitting Rule Discovery Model for Enhancement of Government's Tobacco Control Systems" (2010). *PACIS 2010 Proceedings*. 26.
http://aisel.aisnet.org/pacis2010/26

SMOKERS' CHARACTERISTICS AND CLUSTER BASED QUITTING RULE DISCOVERY MODEL FOR ENHANCEMENT OF GOVERNMENT'S TOBACCO CONTROL SYSTEMS

Shamsul Huda, John Yearwood, Centre for Informatics and Applied Optimization (CIAO), University of Ballarat, Australia, s.huda@ballarat.edu.au, j.yearwood@ballarat.edu.au

Ron Borland, VicHealth Centre for Tobacco Control, The Cancer Council Victoria, Melbourne, Australia, Ron.Borland@cancervic.org.au

Abstract:

Discovery of cluster characteristics and interesting rules describing smokers' clusters and the behavioural patterns of smokers' quitting intentions is an important task in the development of an effective tobacco control systems. In this paper, we attempt to determine the characteristics of smokers' clusters and simplified rule for predicting smokers' quitting behaviour that can provide feedback to build a scientific evidence-based adaptive tobacco control systems. Standard clustering algorithm groups the data based on there inherent pattern. However, they seldom provide human understandable easy description of the clusters'. Again, standard decision tree (SDT) based rule discovery depends on decision boundaries in the feature space. This may limit the ability of SDT to learn intermediate concepts for high dimensional large datasets such as tobacco control. In this paper, we propose a cluster-based rule discovery model (CRDM) that builds conceptual groups from which a set of decision trees (a decision forest) are constructed to find smokers' quitting rules. We also employ a re-labelling of unsupervised cluster (RLUC) approach to determine the characteristics of the clusters. RLUC approach uses re-labelling and decision tree approach to find the characteristics of the smokers' clusters. Experimental results on the tobacco control data set show that decision rules from the decision forest constructed by CRDM are simpler and can predict smokers' quitting intention more accurately than a single decision tree. RLUC approach finds text-based characteristics of the smokers' clusters which are easily understandable for policy makers in the tobacco control systems.

Keywords: Tobacco control systems, Smokers' quitting rule, Univariate Decision Tree, Multivariate decision tree, rule discovery, Smokers' Cluster characteristics.

1 INTRODUCTION:

Tobacco smoking has a large influence on health and is a significant cause of death. It is one of the main causes of death (ITCEP, WHO, 2008) and currently 5.4 million people die every year due to tobacco smoking (ITCEP, WHO, 2008) in the world. Smoking is the top major cause of death (DHS, Melbourne, 2005) and every year 4000 people die due to smoking in Victoria (VTCS, 2008) with an additional cost over $5 billion each year for Victorians (VTCS, 2008). There were approximately 2.9 million people aged 14 yeasr or older who smoked tobacco daily (NDSH, 2008) in Australia, in 2007. In Victoria, 17.3% of adults are regular smokers (CCV, 2004), (Germain, D, 2008). Tobacco smoking causes death for more than 15,000 Australians every year (Begg S, 2003), (Collins D, 2008). Therefore controlling tobacco smoking has become a social demand. However controlling tobacco smoking and determining corresponding policies is a difficult task since it is related to human habit, behaviour and activities as well as relationships with tobacco industries. Therefore policy makers in the Government tobacco control systems need feedback from research to adopt more fruitful policies. This feedback is usually obtained from ground level surveys from smoking population. However survey data does not directly help much and can not explore the overall picture of the effect of the policies.

Researchers in tobacco control systems have used standard decision trees (SDT) in determining smokers' quitting intention to evaluate and develop effective tobacco control policy. Previous attempt on tobacco control using decision tree based approaches were based on pre-defined concept oriented datasets (X., J., Ding) such as demographic, psychological or particular age group (more on these concepts has been described in section-2). However, the separation of smokers' survey data based on predefined groups may bias the generated rule set which may fail to reflect the actual effect of a tobacco control policy. Another approach considered a single standard decision tree (SDT) (J. R. Quinlan, 1987) to generalize the input and target attributes relationship (X., J., Ding, 2008). SDTs show very good expressive power, however, SDTs are univariate (L, Rokach, 2005). This may limit the capacity of an SDT to learn the intermediate concepts for high dimensional large datasets such as tobacco control. Researchers have tried to use multivariate splitting criteria (L, Rokach, 2005) in decision trees. However finding the best multivariate criteria for a high dimensional data set is also complicated and computationally expensive. Therefore, in this paper, we propose a cluster-based rule discovery model (CRDM) for the generation of more compact and simplified rules for enhancement of tobacco control policy. The cluster-based approach builds conceptual groups from which a set of decision trees (a decision forest) is constructed. Then interesting rule sets can be extracted from the decision forest. Extracted rules from the decision forest are simple and show more correct prediction capability.

It is apparent from qualitative analysis of smokers' reactions to various campaigns that some sub-groups of smokers seem to react in quite different ways (Carter S, Borland R, 2001) for the existing tobacco control policies. If these groups could be characterised and the influences on them could be quantified, it might be possible to tailor interventions to be effective with more smokers. Therefore, determination of smokers' clusters and characteristics of those clusters' is an important task in construction of effective tobacco control policy. However, to the best of our knowledge, this type of analysis of tobacco control system has not been done in current tobacco control literature. In this paper, we propose a Re-Labelling of Unsupervised Cluster (RLUC) approach to determine the cluster description where the samples in the smoking population are labelled according to their cluster label. Then re-labelled samples of the smoking population are fed to the decision tree induction algorithm (J. R. Quinlan, 1987) to generate simple text-based characteristics of the smokers' clusters those are easily understandable to the policy makers. The rest of the paper is organized as follows. The next section describes the tobacco control data sets. Section 3 gives a detailed description of the proposed CRDM and cluster characteristics determination approaches. Experimental analysis and results are described in section 4. The conclusion of this study is given in the last section.

2 DATA SETS

International Tobacco Control Policy Evaluation Project (ITC Project) completed a four country survey (known as ITC-4 data) [5], [6] with a target of estimating the impact of psychological and behavioural impact of the key policies of Framework Convention on Tobacco Control (FCTC))[5], [6], [7] organised by the World Health Organization (WHO). The Four-Country Survey was made among randomly selected smokers in four English-speaking countries: Canada, the United States, the United Kingdom, and Australia. The survey consists of four waves. First survey (Wave-1) was completed during October-December 2002. Following every 8 or 9 months (approximately) a survey was conducted. Wave-2 was conducted during May-August 2003, Wave-3 during June–December 2004, and Wave 4 from September–December 2005. Eighty five or more questions have been considered to evaluate the impact of tobacco control policy measures for different waves. Survey question are mainly based on psychosocial – beliefs about smoking, beliefs about quitting, psychosocial questions such as perceived risk and health worry, smoking behaviour such as total minutes to first cigarette, addictedness to cigarettes), knowledge of health effects/tobacco constituents, socio-demographic questions such income, smokers' reaction and outcome on cessation advice and services, smokers' reactions on warning labels, advertising, monitoring of anti-tobacco campaigns, price/taxation and sources of tobacco, smokers' reactions and effect on smoking restrictions. The main outcome questions is whether the smokers' have made any attempt to stop smoking since they were interviewed last or they have stopped smoking for about 6 months. Questions of ITC-4 are described in the Appendix section.

3 METHODOLOGY:

In this paper we first propose a Re-Labelling of Unsupervised Cluster (RLUC) approach to determine the smokers' clusters and characteristics of those clusters and then propose a cluster-based rule discovery model (CRDM) for determining effective tobacco control policy from those clusters. In RLUC, we employ clustering algorithms to find the smokers' clusters. Then, the smoking population is re-labelled based on their clusters which are then fed to SDT to find the characteristics of the clusters. In CRDM, conceptual groups of smoking population from the first step (RLUC) based on the quitting intention as outcome (without re-labelling) are fed to SDT for each cluster. Then a set of SDTs (a decision forest) is constructed. Decision rules are extracted from the decision forest. The next section discusses the detail of RLUC and CRDM.

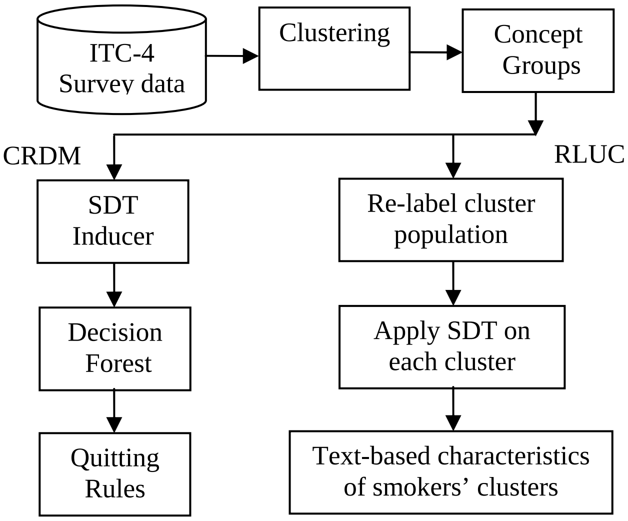


Figure-1: Proposed Cluster characteristics (RLUC) and CRDM for rule discovery in tobacco control

3.1 Standard Decision Tree (SDT)

SDTs (J. R. Quinlan, 1987, L, Rokach, 2005) are one of the popular approaches in predictive data mining tasks and widely used in decision support systems (DSS). A SDT is a rooted tree with a node called the root that has no incoming edge. Nodes with both incoming and outgoing edges are called internal nodes. Nodes with no outgoing edges are called leaves. In general, SDTs are constructed by following a divide and conquer search strategy that recursively partition the training spaces into subspaces according to the value of a single feature. The selection of an input features in partitioning the sample space is done by some goodness measure. The goodness measure ranks the features and the best feature is chosen. Many goodness measure have been proposed such as impurity based criteria (J. R. Quinlan, 1987), likelihood ratio (F. Attneave et.al. ,1959), Gain ratio (J. R. Quinlan, 1987). Each path from a root of the SDT to one of its leaves can easily be transformed to a decision rule by co-joining the intermediate nodes' test-conditions that forms the antecedent part of the rule and class value of the leaf forms the consequent part of the rule.

3.2 Cluster Analysis

Clustering is a process of grouping of a set of samples in a manner that maximizes the intraclass similarity and minimizes the interclass similarity. The process is also known as unsupervised classification where a set of unsupervised data are separated into a discrete set of natural and hidden structures. When sample spaces are clustered the samples within a cluster have high similarity with each other and show high dissimilarity to the samples of other clusters. The clustering process can use various proximity measures (e.g.Minkowski distance (R. Hathaway et.al., 2000), Mahalanobis distance (J. Mao, et.al., 1996), Pearson co-relation (M. Eisen, 1998), Global-K-Means (A.M. Bagirov,et.al., 2008) and various criterion functions (sum square error (A. Likas, et. al. ,2003), Maximum Likelihood (G. McLachlan, et. al., 1997)) in grouping the unsupervised data.

3.3 Proposed Cluster characteristics determination (RLUC) and Cluster-based Rule Discovery Model (CRDM) approaches

To extract the smokers' clusters characteristics from unsupervised classification of the smoking population is a not an easy task. Very few conceptual clustering techniques provide a cluster description along with the clusters. Clustering techniques such as COBWEB (D.H. Fisher, 1987) and CLUSTER (Charles A. Bouman) provide cluster descriptions in some probabilistic measures. However these estimates of clusters are not sufficiently simple to provide effective feedback to tobacco control systems. We have proposed a Re-Labelling of Unsupervised Cluster (RLUC) approach to determine the cluster description. In the RLUC approach, the smoking population form the tobacco survey data are clustered first using a clustering algorithm, Global-K-Means (A.M. Bagirov,et.al., 2008). Global-K-Means is based-on K-means clustering algorithm. It is an incremental algorithm that dynamically adds one cluster centre at a time and uses each data point as a candidate for the k-th cluster centre (A.M. Bagirov,et.al., 2008). A starting point for the k-th cluster centre in this algorithm is computed by minimizing an auxiliary cluster function (A.M. Bagirov,et.al., 2008). Then clustered data from each cluster are re-labelled according to their cluster label. Then re-labelled data of each cluster of the smoking population as well as the results of clustering techniques are fed to the decision tree induction algorithm (EC4.5)(S. Ruggieri, 2002) to generate decision rules for the clusters. The decision rules from the decision trees can describe the cluster characteristics and provides easier interpretation for smokers' clusters for the tobacco control policy makers. The procedure has been presented in Figure-1.

In the second step, we employ a cluster-based rule discovery model (CRDM). In general, SDT is univariate where the decision boundaries in the feature space are geometrically orthogonal to the axis of the feature of a particular decision node. This may limit SDT's ability to learn the intermediate concepts for high dimensional data set. Use of multivariate splitting criteria can overcome the problem up to a certain extent. Multivariate splitting criteria is based on a linear combination of input attributes. However finding best multivariate criteria at each intermediate node of the tree is also complicated and computationally expensive. We propose a cluster-based rule discovery model

(CRDM). In CRDM, the clustered smoking population (without re-labelled, and quitting intention as the outcome variable) from RLUC step is used to obtain the natural groups from the sample space based on their hidden pattern of the smokers. Then an efficient decision tree inducer, an efficient version of C4.5 (EC4.5) (S. Ruggieri, 2002) is applied on the each individual group from which a set of decision trees (a decision forest) is constructed. The decision forest provides cluster based decision rule for smokers' quitting intention. Since several conceptual groups are built based on the hidden data structure, the decision trees based on the conceptual groups become more simple and compact. The CRDM is also presented in Figure-1.

4 EXPERIMENTAL ANALYSIS

4.1 Settings

ITC-4 (Wave-1) G. T. Fong, et.al., (2005) data set has been used for smokers' cluster characteristics and quitting rule discovery in tobacco control and to test the efficiency of the rules. We applied Global-K-Means (A.M. Bagirov,et.al., 2008) to cluster the data sets. Then EC4.5 (S. Ruggieri, 2002) was applied on each cluster to build the decision trees.

4.2 Training Data and Test Data:

In RLUC approach we apply Global K-Means to cluster the smokers' survey data. For CRDM, for each cluster form RLUC, we randomly divide the data into training and test sets. Two-third of data of each cluster has been taken for training and the remaining one-third has been taken as test data. The random division of data into training and test has been done five times. Therefore, five different experiments from the random divisions of each cluster data have been performed for each cluster. The test with lowest error has been adopted for the decision tree. The decision tree constructed from each cluster is applied on the corresponding test set of each cluster to verify the prediction ability of the decision rules. In a set of separate experiments (total five), the whole data set has been divided into training and test sets. EC4.5 (S. Ruggieri, 2002) has been applied on the whole training set and the corresponding SDT is applied on the test set. Finally, the average error rate from all clusters (by CRDM) has been compared with the average error rate of a complete test (a single decision tree without clustering).

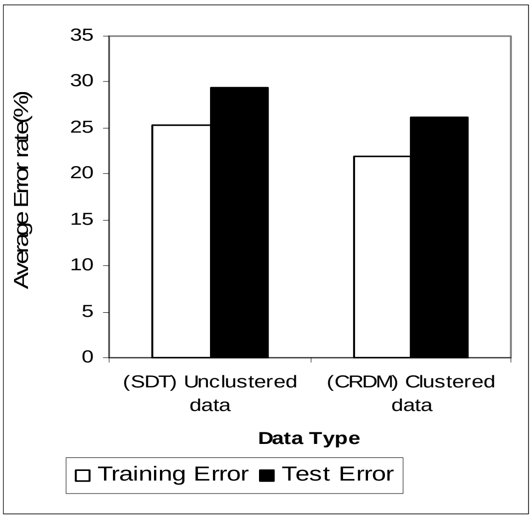


Figure-2: Comparison of average error rate (%) of CRDM (Clustered data) and SDT (Unclustered data)

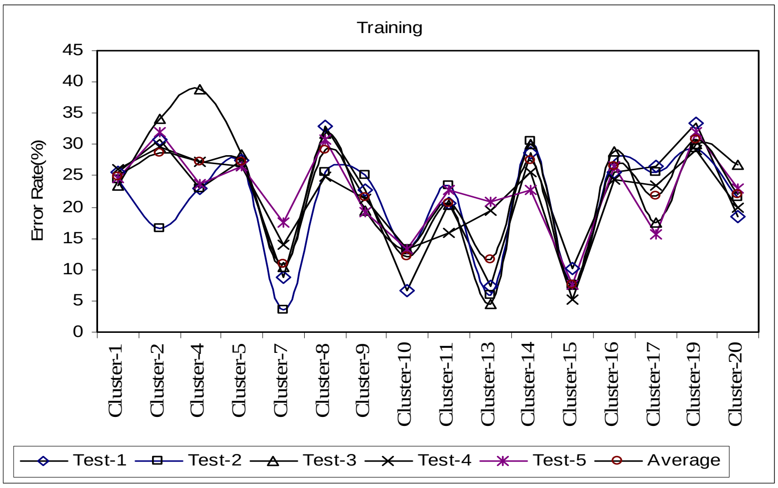


Figure-3: Error rate (%) of five tests and average error rate (%) of CRDM in cluster 1 to 20 in training data.

4.3 Results:

Global K-means finds total 20 clusters. The results have been presented in the Figure-2, 3, 4. Figure-3, 4 describes the error rate of all five tests for cluster-1 to cluster-20 and their average error rate for training and test data. It is seen in figures 2, 3 and 4 that the average error rate (26.21%) by the decision forest of the clustered data obtained by CRDM is less than the error rate (29.35%) by the single SDT of un-clustered data. This proves the effectiveness of CRDM based rule discovery approach. The extracted rules from the decision forest constructed by CRDM have been presented in the Table-1. The attributes' description of the rule is given in the Appendix section. It is seen in Table-1 that rules extracted by CRDM are also very simple. Some of the rules even have only one decision node in CRDM. Clusters-3, 6, 12, 18 have only one member. Therefore these have not been included in the results and less significant. Decision trees constructed by CRDM have small size and the tree-depth is very low. In contrast, the SDT from un-clustered data has 389 nodes and maximum depth 10. This gives very complex rules in which are not suitable for policy makers. Smokers' clusters characteristics by RLUC approach have been presented in the Table-2 and explained in section-4.4. The characteristics of smokers' clusters by RLUC approach are also easy interpretable.

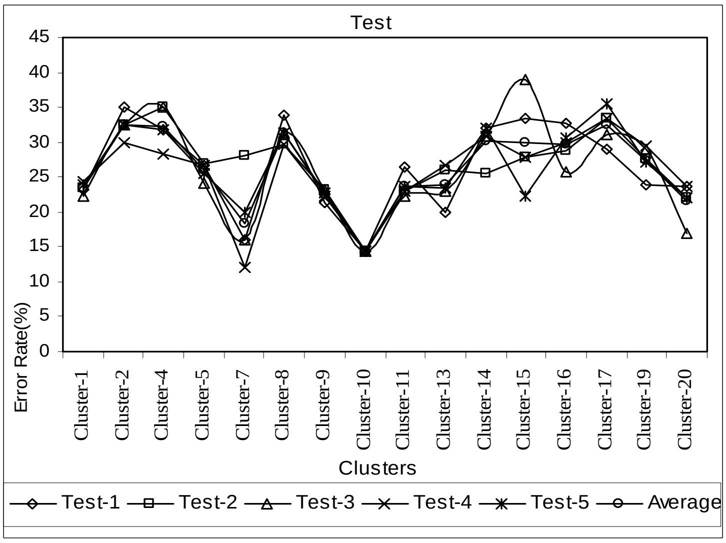
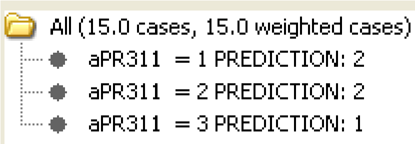


Figure-4: Error rate (%) of five tests and average error rate (%) of CRDM in cluster 1 to 20 of test data.

4.4 Interpretation of extracted rules and cluster characteristics

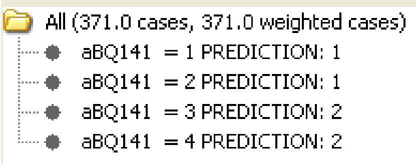
Rules can be easily extracted from the decision forest constructed by CRDM. For a particular decision tree from a cluster (Cluster-10) given below, the rule is extracted for cluster-10 as follows:



```
All (15.0 cases, 15.0 weighted cases)
 ├ aPR311  = 1 PREDICTION: 2
 ├ aPR311  = 2 PREDICTION: 2
 └ aPR311  = 3 PREDICTION: 1
```

In cluster-10, the characteristics of the cluster (in Table-2) is that smokers have (aSB012v>780) (aSB012v = total minutes to first cigarette in the Appendix section).
***The rule: [IF (aPR311 = 3) THEN Smokers made a quit attempt].* Otherwise they did not make a quit attempt .**

Where aPR311 means the question to the smokers as: ***To what extent, if at all, has smoking damaged your health? Choice of answer: 01 – Not at all, 02 – Just a little, 03 – A fair amount, 04 – A great deal.***
Another example of a decision tree for cluster-19 is as follows:



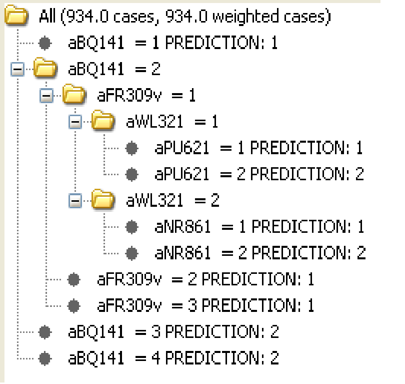
```
All (371.0 cases, 371.0 weighted cases)
 ├ aBQ141  = 1 PREDICTION: 1
 ├ aBQ141  = 2 PREDICTION: 1
 ├ aBQ141  = 3 PREDICTION: 2
 └ aBQ141  = 4 PREDICTION: 2
```

In Cluster-19, the smokers have characteristics (in Table-2) such that ***(aDE212v = 1 AND aSB012v <= 35 AND apu555 <= 31.78)***. This means that smokers have least income (less than 10,000$) and total minutes to first cigarette is less than 35 and price/unit cigarette is greater than 33.

The rule: [IF (aBQ141 =1 OR aBQ141 = 2) THEN Smokers made a quit attempt].

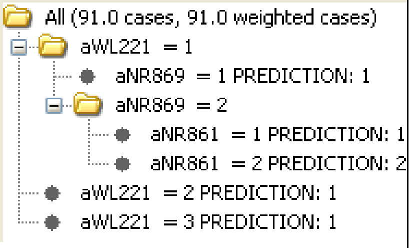
Where aBQ141 means the question to the smokers as: Are you planning to quit smoking? Choice of answer: 01 – Within the next month?, 02 – Within the next 6 months?, 03 – Sometime in the future, beyond 6 months,04 – Not planning to quit. All decision tress have been given in Table-1 from which rules can be extracted, the clusters' characteristics have been given in Table-2 and the attributes meaning have been described in the Appendix section.

Table-1: Extracted rules from decision forest obtained by CRDM (proposed)



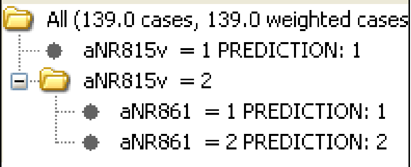
```
All (934.0 cases, 934.0 weighted cases)
 ├ aBQ141  = 1 PREDICTION: 1
 ⊟ aBQ141  = 2
 │  ⊟ aFR309v  = 1
 │  │  ⊟ aWL321  = 1
 │  │  │  ├ aPU621  = 1 PREDICTION: 1
 │  │  │  └ aPU621  = 2 PREDICTION: 2
 │  │  ⊟ aWL321  = 2
 │  │     ├ aNR861  = 1 PREDICTION: 1
 │  │     └ aNR861  = 2 PREDICTION: 2
 │  ├ aFR309v  = 2 PREDICTION: 1
 │  └ aFR309v  = 3 PREDICTION: 1
 ├ aBQ141  = 3 PREDICTION: 2
 └ aBQ141  = 4 PREDICTION: 2
```

Cluster-1



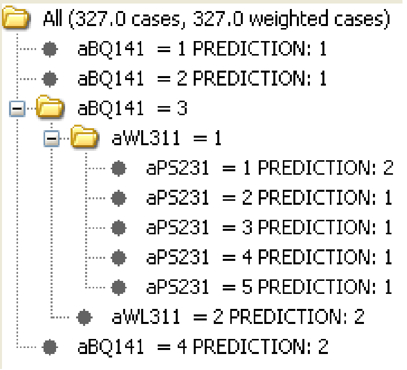
```
All (91.0 cases, 91.0 weighted cases)
 ⊟ aWL221  = 1
 │  ├ aNR869  = 1 PREDICTION: 1
 │  ⊟ aNR869  = 2
 │     ├ aNR861  = 1 PREDICTION: 1
 │     └ aNR861  = 2 PREDICTION: 2
 ├ aWL221  = 2 PREDICTION: 1
 └ aWL221  = 3 PREDICTION: 1
```

Cluster-2



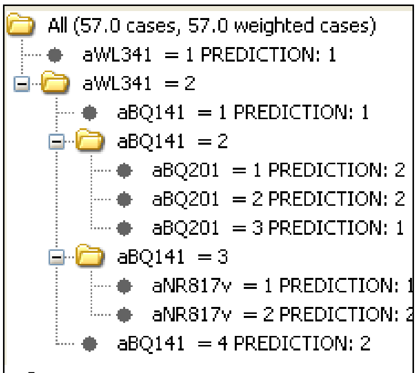
```
All (139.0 cases, 139.0 weighted cases
 ├ aNR815v  = 1 PREDICTION: 1
 ⊟ aNR815v  = 2
    ├ aNR861  = 1 PREDICTION: 1
    └ aNR861  = 2 PREDICTION: 2
```

Cluster-4



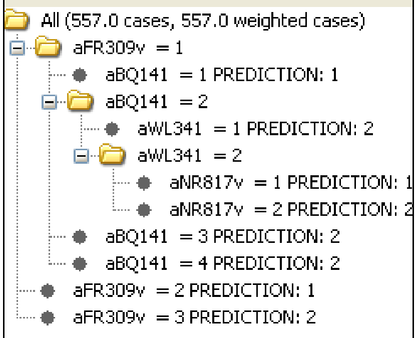
```
All (327.0 cases, 327.0 weighted cases)
 ├ aBQ141  = 1 PREDICTION: 1
 ├ aBQ141  = 2 PREDICTION: 1
 ⊟ aBQ141  = 3
 │  ⊟ aWL311  = 1
 │  │  ├ aPS231  = 1 PREDICTION: 2
 │  │  ├ aPS231  = 2 PREDICTION: 1
 │  │  ├ aPS231  = 3 PREDICTION: 1
 │  │  ├ aPS231  = 4 PREDICTION: 1
 │  │  └ aPS231  = 5 PREDICTION: 1
 │  └ aWL311  = 2 PREDICTION: 2
 └ aBQ141  = 4 PREDICTION: 2
```

Cluster-5



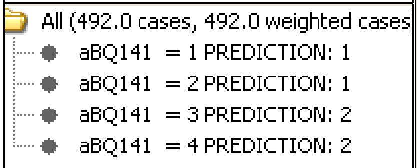
```
All (57.0 cases, 57.0 weighted cases)
  aWL341 = 1 PREDICTION: 1
  aWL341 = 2
    aBQ141 = 1 PREDICTION: 1
    aBQ141 = 2
      aBQ201 = 1 PREDICTION: 2
      aBQ201 = 2 PREDICTION: 2
      aBQ201 = 3 PREDICTION: 1
    aBQ141 = 3
      aNR817v = 1 PREDICTION: 1
      aNR817v = 2 PREDICTION: 2
    aBQ141 = 4 PREDICTION: 2
```

Cluster-7



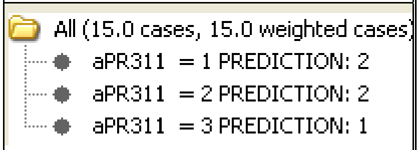
```
All (557.0 cases, 557.0 weighted cases)
  aFR309v = 1
    aBQ141 = 1 PREDICTION: 1
    aBQ141 = 2
      aWL341 = 1 PREDICTION: 2
      aWL341 = 2
        aNR817v = 1 PREDICTION: 1
        aNR817v = 2 PREDICTION: 2
    aBQ141 = 3 PREDICTION: 2
    aBQ141 = 4 PREDICTION: 2
  aFR309v = 2 PREDICTION: 1
  aFR309v = 3 PREDICTION: 2
```

Cluster-8



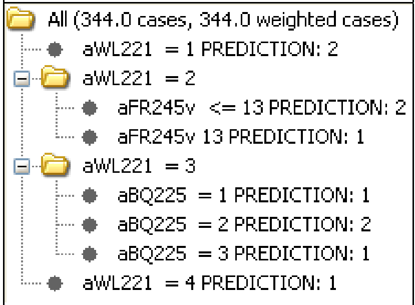
```
All (492.0 cases, 492.0 weighted cases
  aBQ141 = 1 PREDICTION: 1
  aBQ141 = 2 PREDICTION: 1
  aBQ141 = 3 PREDICTION: 2
  aBQ141 = 4 PREDICTION: 2
```

Cluster-9



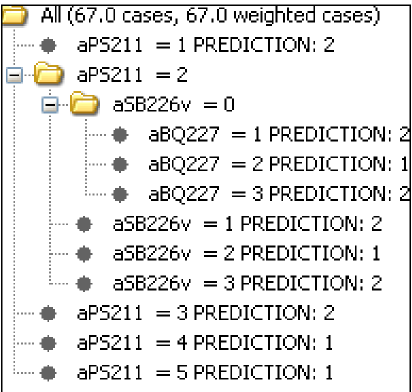
```
All (15.0 cases, 15.0 weighted cases)
  aPR311 = 1 PREDICTION: 2
  aPR311 = 2 PREDICTION: 2
  aPR311 = 3 PREDICTION: 1
```

Cluster-10



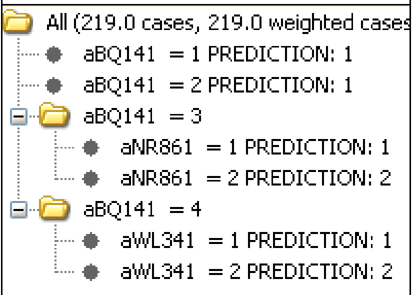
```
All (344.0 cases, 344.0 weighted cases)
  aWL221 = 1 PREDICTION: 2
  aWL221 = 2
    aFR245v <= 13 PREDICTION: 2
    aFR245v 13 PREDICTION: 1
  aWL221 = 3
    aBQ225 = 1 PREDICTION: 1
    aBQ225 = 2 PREDICTION: 2
    aBQ225 = 3 PREDICTION: 1
  aWL221 = 4 PREDICTION: 1
```

Cluster-11



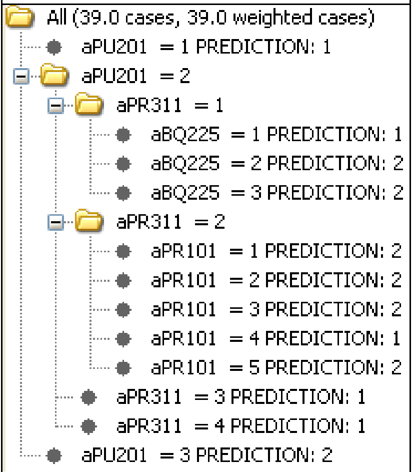
```
All (67.0 cases, 67.0 weighted cases)
  aPS211 = 1 PREDICTION: 2
  aPS211 = 2
    aSB226v = 0
      aBQ227 = 1 PREDICTION: 2
      aBQ227 = 2 PREDICTION: 1
      aBQ227 = 3 PREDICTION: 2
    aSB226v = 1 PREDICTION: 2
    aSB226v = 2 PREDICTION: 1
    aSB226v = 3 PREDICTION: 2
  aPS211 = 3 PREDICTION: 2
  aPS211 = 4 PREDICTION: 1
  aPS211 = 5 PREDICTION: 1
```

Cluster-13



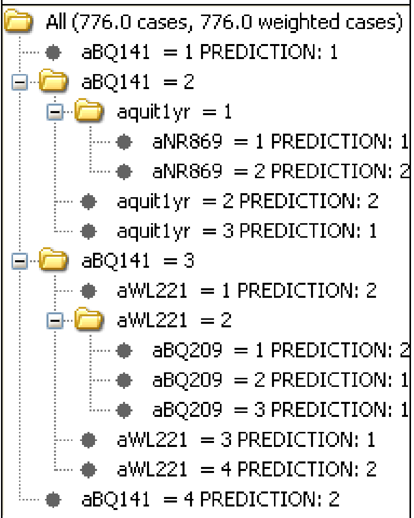
```
All (219.0 cases, 219.0 weighted cases
  aBQ141 = 1 PREDICTION: 1
  aBQ141 = 2 PREDICTION: 1
  aBQ141 = 3
    aNR861 = 1 PREDICTION: 1
    aNR861 = 2 PREDICTION: 2
  aBQ141 = 4
    aWL341 = 1 PREDICTION: 1
    aWL341 = 2 PREDICTION: 2
```

Cluster-14



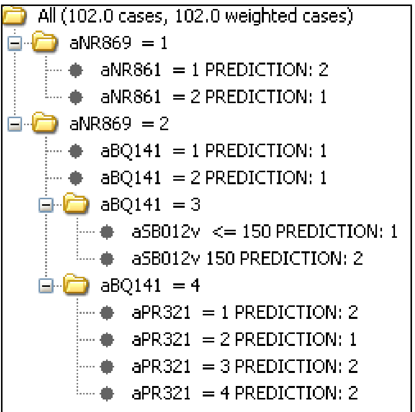
```
All (39.0 cases, 39.0 weighted cases)
  aPU201 = 1 PREDICTION: 1
  aPU201 = 2
    aPR311 = 1
      aBQ225 = 1 PREDICTION: 1
      aBQ225 = 2 PREDICTION: 2
      aBQ225 = 3 PREDICTION: 2
    aPR311 = 2
      aPR101 = 1 PREDICTION: 2
      aPR101 = 2 PREDICTION: 2
      aPR101 = 3 PREDICTION: 2
      aPR101 = 4 PREDICTION: 1
      aPR101 = 5 PREDICTION: 2
    aPR311 = 3 PREDICTION: 1
    aPR311 = 4 PREDICTION: 1
  aPU201 = 3 PREDICTION: 2
```

Cluster-15



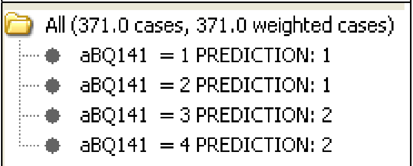
```
All (776.0 cases, 776.0 weighted cases)
  aBQ141 = 1 PREDICTION: 1
  aBQ141 = 2
    aquit1yr = 1
      aNR869 = 1 PREDICTION: 1
      aNR869 = 2 PREDICTION: 2
    aquit1yr = 2 PREDICTION: 2
    aquit1yr = 3 PREDICTION: 1
  aBQ141 = 3
    aWL221 = 1 PREDICTION: 2
    aWL221 = 2
      aBQ209 = 1 PREDICTION: 2
      aBQ209 = 2 PREDICTION: 1
      aBQ209 = 3 PREDICTION: 1
    aWL221 = 3 PREDICTION: 1
    aWL221 = 4 PREDICTION: 2
  aBQ141 = 4 PREDICTION: 2
```

Cluster-16



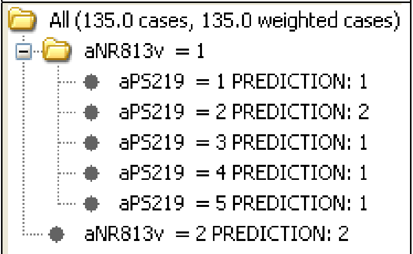
```
All (102.0 cases, 102.0 weighted cases)
  aNR869 = 1
    aNR861 = 1 PREDICTION: 2
    aNR861 = 2 PREDICTION: 1
  aNR869 = 2
    aBQ141 = 1 PREDICTION: 1
    aBQ141 = 2 PREDICTION: 1
    aBQ141 = 3
      aSB012v <= 150 PREDICTION: 1
      aSB012v 150 PREDICTION: 2
    aBQ141 = 4
      aPR321 = 1 PREDICTION: 2
      aPR321 = 2 PREDICTION: 1
      aPR321 = 3 PREDICTION: 2
      aPR321 = 4 PREDICTION: 2
```

Cluster-17



```
All (371.0 cases, 371.0 weighted cases)
  aBQ141 = 1 PREDICTION: 1
  aBQ141 = 2 PREDICTION: 1
  aBQ141 = 3 PREDICTION: 2
  aBQ141 = 4 PREDICTION: 2
```

Cluster-19



```
All (135.0 cases, 135.0 weighted cases)
  aNR813v = 1
    aPS219 = 1 PREDICTION: 1
    aPS219 = 2 PREDICTION: 2
    aPS219 = 3 PREDICTION: 1
    aPS219 = 4 PREDICTION: 1
    aPS219 = 5 PREDICTION: 1
  aNR813v = 2 PREDICTION: 2
```

Cluster-20

Table 2: Cluster characteristics of smokers' by RLUC approach

Cluster-1
aSB012v <= 35 AND aDE212v in {9, 3, 2}
AND apu555 <= 33.5
AND aPS215 in {2, 4, 5}
AND aPS217 in {2, 5, 4, 3}
AND aPS227 in {4, 2, 5, 3}
AND aPS233 in {2, 4}
	Cluster-8
	aSB012v > 35
	AND aSB012v <= 85
	AND aDE212v in {9, 3, 2}
	AND apu555 <= 51
		Cluster-13
		aSB012v > 480
		AND aSB012v <= 600
			Cluster-16
			aSB012v <= 35 AND apu555 <= 33.5 AND aDE212v in {9, 3, 2}
			AND aPS215 in {3, 1}
			AND aPS219 in {5, 1, 3}
			AND aPS229 = 1
			OR aSB012v <= 35
			AND apu555 <= 33.5
			AND aDE212v in {9, 3, 2} AND aPS219 in {5, 1, 3}
			AND aPS229 = 1 AND aPS233 in {1, 3, 5}
			OR
			aSB012v <= 35 AND apu555 <= 33.5 AND aDE212v in {9, 3, 2}
			AND aPS219 in {5, 1, 3} AND aPS229 = 1
			AND aPS233 in {1, 3, 5}

	Cluster-9 aSB012v <= 35 AND apu555 <= 31.78 AND aDE212v = 1 AND age > 45 AND aPS219 in {5, 2, 4, 3} AND aPS231 in {2, 4, 5, 3} AND aPS233 in {2, 4, 5}	Cluster-14 aSB012v > 35 AND aSB012v <= 90 AND aDE212v = 1	
Cluster-2 aSB012v > 600 AND aSB012v <= 780		Cluster-15 aSB012v > 300 AND aSB012v <= 360	
Cluster-4 aSB012v > 210 AND aSB012v <= 300	Cluster-10 aSB012v > 780	Cluster-17 aSB012v > 120 AND aSB012v <= 210	
Cluster-5 aSB012v > 85 AND AND aSB012v <= 120 AND aDE212v in {9, 3, 2}		Cluster-19 aSB012v <= 35 AND apu555 <= 31.78 AND aDE212v = 1	
Cluster-7 aSB012v > 360 AND aSB012v <= 480	Cluster-11 aSB012v <= 85 AND apu555 > 51 AND aDE212v in {9, 3, 2}	Cluster-20 aSB012v <= 35 AND apu555 > 37 AND aDE212v = 1	

5 CONCLUSIONS:

A cluster-based rule discovery model (CRDM) and a Re-labelling of Unsupervised Classification (RLUC) for predicting smokers' quitting intentions and for determination of smokers' cluster characteristics have been proposed for tobacco control systems that helps tobacco control policy makers in determining scientific-evidence based, cost-effective and adaptive policy. The cluster-based approach in CRDM is able to overcome the univariate problem of SDT for high dimensional data (such as tobacco control data). Experimental analysis on the real tobacco control data set shows that CRDM generates more simplified and compact decision rules than a single SDT for tobacco control system. In a single SDT, the tree size becomes large and the depth of the tree is high which means that very complex rules are generated. Moreover, the average prediction error rate of CRDM is less than a single SDT. The RLUC approach provides simple text-based easily recognizable characteristics of smokers' cluster that would help identify the less affected smokers' group by the current regulations in tobacco control systems. The results have been analyzed by the experts in Cancer Council, Melbourne and found to be useful for Tobacco Control Systems. In future we will apply CRDM and RLUC on the other waves of ITC-4 survey of Tobacco Control Systems.

References:

ITCEP,International Tobacco Control policy Evaluation Project (ITCEP), http://www.itcproject.org/.
WHO, 2008, WHO REPORT on the global TOBA CCO epidemic, 2008, The MPOWER package

http://www.who.int/tobacco/mpower/gtcr_download/en/index.html.

DHS, 2005, Department of Human Services 2005, Victorian Burden of Disease Study, DHS, Melbourne.

NDSH, 2008, 2007 National Drug Strategy Household Survey, First results, April 2008, Australian Institute of Health and Welfare, Canberra, Cat. no. PHE 98, http://www.aihw.gov.au/publications/phe/ndshs07-fr/ndshs07-fr-no-questionnaire.pdf

VTCS, 2008, Victorian Tobacco Control Strategy, 2008–2013, Report produced by Published by the Victorian Government Department of Human Services Melbourne, Victoria, Copyright State of Victoria 2008, http://www.health.vic.gov.au/tobaccoreforms/downloads/vtcs0813.pdf

CCV, 2004, The Cancer Council Victoria 2004, Centre for Behavioural Research in Cancer, http://www.cancervic.org.au..

Germain, D., Wakefi eld, M. & Durkin, S. (2008) 2008 Smoking prevalence and consumption in Victoria: Key findings from 1998–2007 population surveys, Centre for Behavioural Research in Cancer, The Cancer Council of Victoria.

Begg S, Vos T, Barker B, Stevenson C, Stanley L, Lopez A. The burden of disease and injury in Australia 2003. PHE 82. Canberra: Australian Institute for Health and Welfare-2007. http://www.aihw.gov.au/publications/index.cfm/title/10317

Collins D, Lapsley H. The costs of tobacco, alcohol and illicit drug abuse to Australian society in 2004–05. P3 2625. Canberra: Department of Health and Ageing; 2008. http://www.nationaldrugstrategy.gov.au/internet/drugstrategy/publishing.nsf/Content/mono64/$File/mono64.pdf

X., J., Ding, S., Bedingfield et al , "A Decision Tree Approach for Predicting Smokers' Quit Intentions", Journal of Electronic Science and Technology of China, vol-6(3), pp 220-224.

J. R. Quinlan, "C4.5 programs for machine learning", Moragn Kaufmann, 1987

X., J., Ding, S.,Bedingfield et al, "A Rule Based Approach to Modeling the Effect of Tobacco Control Policies" 4th International conference on Information and Automation for sustainability(ICIAES), 2008, pp 496-501.

L, Rokach, O., Maimon, "Top-Down induction of decision trees classifier- A survey" , IEEE transaction on system, man and cybernetics Part-C, Vol-35(4), 2005, pp 478-487.

G. T. Fong, K. M. Cummings, R. Borland, G. B. Hastings, P. Hyland, G. A. Giovino, D. Hammond, and M. E. Thompson, "The conceptual framework of the International Tobacco Control (ITC), Policy Evaluation Project," Tobacco Control, vol. 15, Suppl. 3, pp. 3-11, 2005.

M. E. Thompson, G. T. Fong, D. Hammond, C. Boudreau, P. Driezen, P. Hyland, R. Borland, K. M. Cummings, G. B. Hastings, M. Siahpush, A. M. Machintosh, and F. L. Laux, "Methods of the International Tobacco Control (ITC) Four Country Survey", Tobacco Control, vol. 15, Suppl. 3, pp. 12-18, 2006.

A. Hyland, R. Borland, Q. Li, H-H. Yong, A. McNeill, G. T. Fong, R. J. O'Connor, and K. M. Cummings, "Individual- level predictors of cessation behaviours among participants in the International Tobacco Control (ITC) Four Country Survey", Tobacco Control, vol. 15, Suppl. 3, pp. 83-94, 2006.

J. R. Quinlan, "Simplifying decision trees," Int. J. Man-Mach. Studies, vol. 27, pp. 221–234, 1987.

F. Attneave, Applications of Information Theory to Psychology. New York: Holt, Rinehart and Winston, 1959.

R. Hathaway, J. Bezdek, and Y. Hu, "Generalized fuzzy c-means clustering strategies using L norm distances," IEEE Trans. Fuzzy Syst., vol. 8, no. 5, pp. 576–582, Oct. 2000.

J. Mao and A. Jain, "A self-organizing network for hyperellipsoidal clustering (HEC)," IEEE Trans. Neural Netw., vol. 7, no. 1, pp. 16–29, Jan. 1996.

M. Eisen, P. Spellman, P. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," in Proc. Nat. Acad. Sci.USA, vol. 95, 1998, pp. 14 863–14 868.

A.M. Bagirov, Modified global k-means algorithm for minimum sum-of-squares clustering problems, Pattern Recognition, Volume 41, 2008, pp 3192-3199.

A. Likas, N. Vlassis, and J. Verbeek, "The global K-means clustering algorithm," Pattern Recognition, vol. 36, no. 2, pp. 451–461, 2003.

G. McLachlan and T. Krishnan, The EM Algorithm and Extensions. New York: Wiley, 1997.

S. Ruggieri, “Efficient C4.5”, IEEE transaction on knowledge and data engineering, vol-14 (2), 2002, pp 438-444.

Carter S, Borland R and Chapman S., “Finding the strength to kill your best friend - Smokers talk about smoking and quitting.” Sydney: Australian Smoking Cessation Consortium and GlaxoSmithKline Consumer Healthcare, May 2001.

D.H. Fisher, “Knowledge Acquisition via incremental conceptual clustering”, Machine Learning 2, 1987, pp 139-172.

Charles A. Bouman, “CLUSTER: An Unsupervised Algorithm for Modeling Gaussian Mixtures”, School of Electrical Engineering Purdue University, http://www.ece.purdue.edu/~bouman

Appendix: Attributes

Prediction=1 (Quit Attempt)

Prediction=2 (No Quit Attempt)

aBQ141

a. Are you planning to quit smoking:

01 – Within the next month?

02 – Within the next 6 months?

03 – Sometime in the future, beyond 6 months

04 – Not planning to quit

aFR250V

cigarettes per day

0= 1-10cigs, 1= 11-20cigs, 2= 21-30cigs, 3= 31+ cigs

aPU621

In the last 6 months, since, have you spent money on cigarettes that you knew would be better spent on household essentials like food.

01 – YES, 02 – NO

aNR861V

Since [LSD], have you received advice or information about quitting smoking from

Telephone or quit line services?

1-Yes, 2-No, 7- NA, 8- Refused, 9- Don't know

aNR815V

quitting RX from doctor, overall (incl those who did not visit the doctor)

AWL221

In the last month, have the warning labels stopped you from having a cigarette when you were about to smoke one? Would you say:

01 – Never, 02 – Once, 03 – A few times, 04 – Many times

ABQ201

In the past 6 months, have each of the following things led you to think about quitting, not at all, somewhat, or very much:

Concern for your personal health?

01 – Not at all, 02 – Somewhat, 03 – Very much

AFR309V

Smoking status: 1=daily, 2=weekly, 3=monthly, 4=quit<1mth, 5=quit 1-6m, 6=quit>6m

aNR817v

pamphlet on quitting, from doctor, overall (incl those who did not visit the doctor)

aNR869

Since [LSD], have you received advice or information about quitting smoking from. Local stop-smoking services (such as clinics or specialists)?

aBQ225

In the past 6 months, have each of the following things led you to think about quitting, not at all, somewhat, or very much:

Advertisements or information about the health risks of smoking? 01 – Not at all, 02 – Somewhat, 03 – Very much)

aPS215

Please tell me whether you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree with each of the following statements. If you had to do it over again, you would not have started smoking 01 – Strongly agree, 02 – Agree, 03 – Neither agree nor disagree, 04 – Disagree, 05 – Strongly disagree

aNR813v

referral from doctor to help stay quit, overall (incl those who did not visit the doctor)

aFR260v

Derived variable: Heaviness of smoking index

aPS211

Please tell me whether you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree with each of the following statements. You enjoy smoking too much to give it up. 01 – Strongly agree, 02 – Agree, 03 – Neither agree nor disagree, 04 – Disagree, 05 – Strongly disagree,

aET221

Which of the following best describes smoking in your home? (read)

01 – Smoking is allowed anywhere in your home, 02 – Smoking is never allowed anywhere in your home, 03 – Something in between

aSB031

Do you consider yourself addicted to cigarettes? 01 – Not at all, 02 – Yes–somewhat addicted, 03 – Yes–very addicted

aFR245

On average, how many cigarettes do you smoke each day/week, including both factory-made and roll-your own cigarettes?

aBQ209

In the past 6 months, have each of the following things led you to think about quitting, not at all, somewhat, or very much:

The price of cigarettes? 01 – Not at all

02 – Somewhat, 03 – Very much

aPU555

Calculated variable: price per unit, regardless of packaging.

Aquit1yr

Tried to quit in the last year: 1=never, 2=tried>1yr ago, 3=tried within last year

aSB221

In the last month—since **[1M Anchor]**, have you [AUS/UK=stubbed] [CAN/US/=butted] out a cigarette before you finished it because you thought about the harm of smoking?

01 – YES , 02 – NO

aSB012v

total minutes to first cigarette (continuous)

aDE312v

Education 3-High; 2-Medium; 1-Low

aBQ221

In the past 6 months, have each of the following things led you to think about quitting, not at all, somewhat, or very much:

Free or lower-cost stop-smoking medication?

01 – Not at all, 02 – Somewhat, 03 – Very much

aBQ229

Setting an example for children?

aBQ201

Concern for your personal health?

01 – Not at all, 02 – Somewhat, 03 – Very much

aBQ227

Warning labels on cigarette packages? 01 – Not at all, 02 – Somewhat, 03 – Very much**)**

aBQ223

Availability of telephone helpline/Quitline/informat ion line?

01 – Not at all, 02 – Somewhat, 03 – Very much

aBQ203

Concern about the effect of your cigarette smoke on non-smokers? 01 – Not at all,02 – Somewhat, 03 – Very much

aSB226v

a. In the last month— since, have you [AUS/UK=stubbed] [CAN/US/=butted] out a cigarette before you finished it because you thought about the harm of smoking?

b. Was that once, a few times, or lots of times?

01 – Once, 02 – A few times, 03 – Lots of times

aPS223

Smoking is an important part of your life.

Please tell me whether you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree with each of the following statements. 01 – Strongly agree 02 – Agree 03 – Neither agree nor disagree 04 – Disagree

05 – Strongly disagree

aBQ121

How easy or hard would it be for you to completely quit smoking if you wanted to?

01 – Very easy, 02 – Somewhat easy, 03 – Neither easy nor hard, 04 – Somewhat hard

05 – Very hard

aPS229

Please tell me whether you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree with each of the following statements. People who are important to you believe that you should not smoke. 01 – Strongly agree, 02 – Agree, 03 – Neither agree nor disagree, 04 – Disagree, 05 – Strongly disagree

aWL341

In the last month, have you made any effort to avoid looking at or thinking about the warning labels: by not buying packs with particular labels?

01 – YES, 02 – NO

aWL211

In the last month, how often, if at all, have you read or looked closely at the warning labels on cigarette packages?

01 – Never, 02 – Rarely, 03 – Sometimes

04 – Often, 05 – Very often

aPS227

Please tell me whether you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree with each of the following statements. You have strong mixed emotions both for and against smoking, all at the same time.

01 – Strongly agree, 02 – Agree, 03 – Neither agree nor disagree, 04 – Disagree, 05 – Strongly disagree

aSB013v

Minutes to first cigarette

aSB205

The following questions ask you about how often you've had certain thoughts in the last month, that is, For each question, please answer using. Think about the harm your smoking might be doing to you?

01 – Never, 02 – Rarely, 03 – Sometimes

04 – Often, 05 – Very Often

aPS229

People who are important to you believe that you should not smoke.

aDE212v

Income: 1-Low; 2-Medium; 3-High

aPS217

Please tell me whether you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree with each of the following statements. Smoking calms you down when you are stressed or upset. 01 – Strongly agree, 02 – Agree, 03 – Neither agree nor disagree, 04 – Disagree, 05 – Strongly disagree

aPS231

Please tell me about the following statements. There are fewer and fewer places where you feel comfortable about smoking. 01 – Strongly agree, 02 – Agree, 03 – Neither agree nor disagree, 04 – Disagree, 05 – Strongly disagree

aPS233

Please tell me about the following statements. Society disapproves of smoking. . 01 – Strongly agree, 02 – Agree, 03 – Neither agree nor disagree, 04 – Disagree, 05 – Strongly disagree

aPR311

To what extent, if at all, has smoking damaged your health?

01 – Not at all, 02 – Just a little, 03 – A fair amount, 04 – A great deal

aFR309v

Smoking status: 1=daily, 2=weekly, 3=monthly, 4=quit<1mth, 5=quit 1-6m, 6=quit>6m

aNR861

Have you received advice or information about quitting smoking from Telephone or quit line services?, 1-Yes, 2-No,7-NA,8-Refused, 9-Don't know

--In the last month, have you made any effort to avoid looking at or thinking about the warning labels, 01–Yes, 02 – No

aWL311- by covering the warnings up?

aWL321- by keeping the pack out of sight?

aWL331- by using a cigarette case or some other pack?

aWL341- by not buying packs with particular labels?

aPR101

In general, how would you describe your health? Is it, 01–Poor,02–Fair,03–Good,04–Very good,05–Excellent

aPR321

To what extent has smoking lowered your quality of life? Please answer using.01 – Not at all,02 – Just a little,03 – A fair amount,04 – A great deal

aPS219

Please tell me about-You spend too much money on cigarettes. 01 – Strongly agree, 02 – Agree, 03 – Neither agree nor disagree, 04 – Disagree, 05 – Strongly disagree