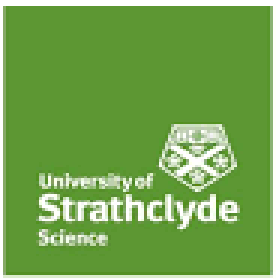


DEPARTMENT OF MATHEMATICS & STATISTICS

UNIVERSITY OF STRATHCLYDE
DEPARTMENT OF MATHEMATICS AND STATISTICS

HARMFUL ALCOHOL CONSUMPTION IN SCOTTISH RURAL AND URBAN CONTEXTS

A QUANTITATIVE APPROACH USING SELF-REPORTED DATA

by

LOLIETT VALDES CASTILLO

Reg. No. 202250524

MSc in Applied Statistics with Data Science

2024

Statement of work in project

The work contained in this project is that of the author and where material from other sources has been incorporated full acknowledgement is made.

Signed: 

Print Name: Loliett Valdes Castillo

Date: 1st August 2024

Supervised by Dr Alison Gray

TABLE OF CONTENT

LIST OF FIGURES ..... **V**

LIST OF TABLES ..... **VI**

1. INTRODUCTION **1**

1.1 BACKGROUND 1

1.2 RESEARCH AIM 5

2. METHODOLOGY **6**

2.1 DATA SOURCE 6

2.2 DATASET 6

2.3 DATA COLLECTION 7

2.4 PRELIMINARY DATA PROCESSING AND VARIABLE SELECTION 8

2.5 OUTPUT AND VARIABLES SELECTED 8

2.6 URBAN AND RURAL DATASETS 13

2.7 HANDLING WITH HIGHLY CORRELATED VARIABLE AND LOW REPRESENTATION VARIABLE. 13

2.8 SPLITTING THE DATASET IN TRAINING AND TEST SUBSETS. 16

2.9 HANDLING THE MISSING DATA 16

2.10 HANDLING THE CLASS IMBALANCE 17

2.10.1 OVERSAMPLING *17*

2.10.2 SMOTE *18*

2.11 METHODS 18

2.11.1 DESCRIPTIVE STATISTICS AND VISUALISATIONS *18*

2.11.2 CHI-SQUARED TEST *19*

2.11.3 LOGISTIC REGRESSION *20*

2.11.4 RANDOM FOREST FOR CLASSIFICATION *23*

3. RESULTS **28**

3.1 CHARACTERISTICS OF THE URBAN & RURAL, URBAN AND RURAL DATASETS 28

3.2 DISTRIBUTION OF THE OUTPUT VARIABLE IN THE URBAN & RURAL, URBAN AND RURAL DATASETS 32

3.3 RESULTING DATASETS AFTER DATA SPLITTING 37

3.4 MISSING DATA ANALYSIS AND RESULTING DATASETS AFTER DELETION 37

3.5 RESULTING DATASETS AFTER OVERSAMPLING 38

3.6 RESULTING DATASETS AFTER SMOTE 39

3.7 CHI-SQUARED 40

3.8 URBAN & RURAL DATASETS 41

3.8.1 LR IN URBAN & RURAL DATASETS *41*

3.8.2 CROSS-VALIDATION IN URBAN & RURAL MODELS *43*

3.8.3 FITTING AND INTERPRETING THE URBAN & RURAL MODEL *44*

3.8.4 EVALUATING THE URBAN & RURAL MODEL IN THE TEST SUBSET *47*
3.9 URBAN DATASETS 48
3.9.1 LR IN URBAN DATASETS *48*
3.9.2 CROSS-VALIDATION IN URBAN MODELS *50*
3.9.3 FITTING AND INTERPRETING THE URBAN MODEL *50*
3.9.4 EVALUATING THE URBAN MODEL IN THE TEST SUBSET *54*
3.10 RURAL DATASETS 55
3.10.1 LR IN RURAL DATASETS *55*
3.10.2 CROSS-VALIDATION IN RURAL MODELS *57*
3.10.3 FITTING AND INTERPRETING THE RURAL MODEL *57*
3.10.4 EVALUATING THE RURAL MODEL IN THE TEST SUBSET *61*
3.11 CONSIDERATIONS ABOUT THE THREE MODELS 62
3.11.1 VARIABLES CONSIDERED IN THE THREE MODELS *62*
3.11.2 RESULT INTERPRETATION IN THE THREE MODELS *62*
3.11.3 EVALUATING THE THREE MODELS IN THE TEST SUBSETS *64*
3.12 RF IN THE RURAL DATASETS 65
4. DISCUSSION 68
4.1 ALCOHOL CONSUMPTION PATTERNS CONCERNING SOCIODEMOGRAPHICS AND SOCIOECONOMICS FACTORS 68
4.2 PREDICTING MODELLING ACROSS THE THREE DATASETS 70
4.3. LIMITATIONS 71
5. CONCLUSION 73
REFERENCES 75
APPENDIX: SUPPLEMENTARY TABLES 81
LIST OF SUPPLEMENTARY TABLES 81
ANNEXES 141
ANNEX A. COPY OF EMAIL RECEIVED FROM UK DATA SERVICE 141
ANNEX B. CONVERSION USED FOR THE SHES TO CONVERT VOLUMES OF ALCOHOL REPORTED IN THE SURVEY INTO UNITS 142

LIST OF FIGURES

FIGURE 3.1 PROPORTION OF SOCIOECONOMIC VARIABLES ACROSS THE THREE DATASETS 29
FIGURE 3.2 PROPORTION OF HEALTH-RELATED VARIABLES ACROSS THE THREE DATASETS 30
FIGURE 3.3 PROPORTION OF EDUCATIONAL AND SOCIOECONOMIC-RELATED VARIABLES ACROSS THE THREE DATASETS 31
FIGURE 3.4 ALCOHOL CONSUMPTION ACROSS THE THREE DATASETS 32
FIGURE 3.5 ALCOHOL CONSUMPTION RELATIVE TO WEEKLY LIMITS BY SOCIODEMOGRAPHIC VARIABLES ACROSS A. URBAN & RURAL DATASET (IN THIS PLOT AS OVERALL DATASET), B. URBAN DATASET AND C. RURAL DATASET 33
FIGURE 3.6 ALCOHOL CONSUMPTION RELATIVE TO WEEKLY LIMITS BY HEALTH-RELATED VARIABLES ACROSS A. URBAN & RURAL DATASET (IN THIS PLOT AS OVERALL DATASET), B. URBAN DATASET AND C. RURAL DATASET 34
FIGURE 3.7 ALCOHOL CONSUMPTION RELATIVE TO WEEKLY LIMITS BY EDUCATIONAL AND SOCIOECONOMIC-RELATED VARIABLES ACROSS A. URBAN & RURAL DATASET (IN THIS PLOT AS OVERALL DATASET), B. URBAN DATASET AND C. RURAL DATASET 35
FIGURE 3.8 ALCOHOL CONSUMPTION BY BEVERAGE TYPE ACROSS A. URBAN & RURAL DATASET (IN THIS PLOT AS OVERALL DATASET), B. URBAN DATASET AND C. RURAL DATASET 36
FIGURE 3.9 MISSING DATA PER VARIABLE 37
FIGURE 3.10 INTERSECTION OF MISSING DATA 37
FIGURE 3.11 STANDARDISED RESIDUALS PLOT 40
FIGURE 3.12 CHI-SQUARED CONTRIBUTION PLOT 40
FIGURE 3.13 URBAN & RURAL MODEL: ODDS RATIOS AND 95% CONFIDENCE INTERVALS 46
FIGURE 3.14 URBAN & RURAL TEST SUBSET: MODEL'S EVALUATION RESULTS 48
FIGURE 3.15 URBAN MODEL: ODDS RATIOS AND 95% CONFIDENCE INTERVALS 53
FIGURE 3.16 URBAN TEST SUBSET: MODEL'S EVALUATION RESULTS 55
FIGURE 3.17 RURAL MODEL: ODDS RATIOS AND 95% CONFIDENCE INTERVALS 60
FIGURE 3.18 RURAL TEST SUBSET: MODEL'S EVALUATION RESULTS 61
FIGURE 3.19 VARIABLE IMPORTANCE FROM RF 67

LIST OF TABLES

TABLE 2.1 VARIABLES PRELIMINARY SELECTED FOR THE ANALYSIS 10
TABLE 2.2 RE-CODING OF UNDER-REPRESENTED CATEGORIES 14
TABLE 2.3 CLASS IMBALANCE IN THE OUTPUT VARIABLE (OVERLIM15) 17

TABLE 2.1 VARIABLES PRELIMINARY SELECTED FOR THE ANALYSIS 10
TABLE 2.2 RE-CODING OF UNDER-REPRESENTED CATEGORIES 14
TABLE 2.3 CLASS IMBALANCE IN THE OUTPUT VARIABLE (OVERLIM15) 17

TABLE 3.1 DIMENSIONS OF TRAINING AND TEST SUBSETS 37
TABLE 3.2 DIMENSIONS OF TRAINING AND TEST SUBSETS BEFORE AND AFTER LISTWISE DELETION 38
TABLE 3.3 DATASETS OBTAINED BY OVERSAMPLING 38
TABLE 3.4 DATASETS OBTAINED BY SMOTE 39
TABLE 3.5 URBAN & RURAL DATASETS: STEPWISE SELECTION RESULTS 42
TABLE 3.6 URBAN & RURAL MODELS: CROSSVALIDATION RESULTS 43
TABLE 3.7 URBAN & RURAL MODEL: PREDICTORS AND THEIR CATEGORIES 44
TABLE 3.8 URBAN & RURAL MODEL: METRICS IN THE TEST SUBSET 47
TABLE 3.9 URBAN DATASETS: STEPWISE SELECTION RESULTS 49
TABLE 3.10 URBAN MODELS: CROSSVALIDATION RESULTS 50
TABLE 3.11 URBAN MODEL: PREDICTORS AND THEIR CATEGORIES 51
TABLE 3.12 URBAN MODEL: METRICS IN THE TEST SUBSET 54
TABLE 3.13 RURAL DATASETS: STEPWISE SELECTION RESULTS 56
TABLE 3.14 RURAL MODELS: CROSSVALIDATION RESULTS 57
TABLE 3.15 RURAL MODEL: PREDITORS AND THEIR CATEGORIES 57
TABLE 3.16 RURAL DATASET: METRICS IN THE TEST SUBSET 61
TABLE 3.17 RURAL DATASETS: RF RESULTS 65
TABLE 3.18 RF: CROSSVALIDATION RESULT 66
TABLE 3.19 RURAL TEST SUBSET: RF METRICS 67

1. INTRODUCTION

From the perspective of toxicology, alcohol is considered a psychoactive substance with dependence-producing properties (NIAAA, Dias da Silva et al., 2021). As a central nervous system depressant, alcohol affects brain function by enhancing the effects of the inhibitory neurotransmitter gamma-aminobutyric acid (GABA) and inhibiting the excitatory neurotransmitter glutamate. This dual action results in the characteristic effects of alcohol consumption, such as impaired coordination, slurred speech, and decreased inhibition, which can lead to risky behaviours. Chronic consumption can lead to physical dependence, characterised by tolerance and withdrawal symptoms, and psychological dependence, where alcohol becomes central to the individual's life (Lappas and Lappas, 2022, Dias da Silva et al., 2021).

1.1 BACKGROUND

Alcohol has been produced and consumed by humans for thousands of years. Initially used for medicinal purposes, its recreational use soon spread. The earliest evidence of alcohol production dates back to around 7000-6600 BCE in Jiahu, China, where residues of a fermented beverage made from rice, honey, and fruit were found. In ancient Egypt, beer and wine were integral to daily life and religious rituals. The Greeks and Romans also highly regarded wine, associating it with their gods and incorporating it into their symposia and feasts (McGovern, 2009, Phillips, 2014).

A notable example of the significant connection of alcoholic beverages with societies and cultures can be appreciated in Scotland, where whisky, or *uisge beatha* ["water of life" in Gaelic], symbolises the Scottish identity and craftsmanship. The earliest written record of whisky distillation in Scotland dates back to 1494, recorded in the Exchequer Rolls. In this historical record, Friar John Cor was instructed by the King to produce *aquavitie*, underscoring the beverage's importance even in medieval times (Daiches, 1978, Martin, 2008).

The social and economic influence of alcoholic drinks grew in the modern era. Establishing pubs and taverns in Europe created communal spaces for social interaction and political discussion (McGovern, 2009, Phillips, 2014). In Scotland, whisky remains until the present vital to the economy and cultural export. Protected by law, Scotch whisky is internationally recognised for its quality and heritage. It contributes significantly to the economy and plays a central role in tourism, attracting visitors worldwide to Scotland's distilleries and whisky trails (Bower, 2016).

However, the negative impacts of excessive alcohol consumption also became a global concern as its use spread. As early as the 19th and 20th centuries, prohibitionist or alcohol control movements emerged in several countries (McGovern, 2009, Phillips, 2014). One of the most well-known took place in the United States (U.S.A.) and ended up implementing the Volstead Act (Blocker, 2006). This act aimed to curb alcohol consumption and its associated problems by banning its production, sale, and transport during what is known as the "Prohibition-era" (1920-1933). However, during this period, the rise of illegal alcohol production and organised crime was also observed. The eventual repeal of Prohibition highlighted the challenges of regulating a substance deeply embedded in social and cultural practices.

Today, alcohol regulation varies worldwide, with many countries implementing policies to control consumption due to its health and social impacts (Esser and Jernigan, 2018, WHO, 2018). The World Health Organization (WHO) (2010, 2023, 2018) and other public health bodies, such as the National Institute on Alcohol Abuse and Alcoholism in the U.S. (NIAAA, 2024), the Alcohol Health Alliance in the UK (AHA, 2023), the Alcohol Focus Scotland (AFS, 2021) and the Scottish Health Action on Alcohol Problems (SHAAP, 2021), continue to address the complex issues surrounding alcohol use and advocate for reducing harmful drinking and its consequences.

According to the WHO's Global Status Report on Alcohol and Health (2018), problematic alcohol consumption is estimated to cause over 3 million deaths annually, representing 5.3% of global deaths. Alcohol consumption accounts for 5.1% of the global burden of disease and injury, measured in disability-adjusted life years (DALYs). Additionally, harmful alcohol consumption is linked to mental and behavioural disorders, non-communicable diseases such as liver cirrhosis and cardiovascular problems, as well as both unintentional and intentional injuries, and over 200 other diseases and injuries globally.

The impact of harmful alcohol consumption extends beyond health, with significant social and economic implications (WHO, 2010). Alcohol abuse can lead to familial disruptions, domestic violence, and child neglect, placing additional strain on social services. The broader community is also affected by public safety concerns, such as traffic accidents and alcohol-fueled violence, which require considerable public expenditure to manage. The increase in alcohol consumption leads to a surge in healthcare expenses due to the rise in alcohol-related illnesses and injuries and a need for law enforcement and judicial resources to address alcohol-induced crimes and accidents. Additionally, it provokes the loss of productivity from workforce absenteeism and unemployment caused by alcohol dependence and related health problems (Collaborators, 2018). The WHO emphasises that these combined social and economic impacts hinder sustainable development and pose significant challenges to societal well-being (WHO, 2023).

In Scotland, 1 in 5 people report problematic alcohol consumption, meaning 22% of the population exceeds the recommended weekly limit of 14 units (ScotCen Social Research, 2022a). The current UK Chief Medical Officers' Low-Risk Drinking Guidelines establish this limit as the threshold above which alcohol consumption is defined as harmful (Department of Health England, 2016). In 2022, alcohol-related deaths reached 1,276 (National Records of Scotland (NRS), 2023), and hospital admissions due to alcohol-related conditions totalled 31,206 (PHS, 2024). It is estimated that between £5-10 billion are dedicated annually to health, social care, crime, productive capacity, and other costs related to alcohol consumption (Bhattacharya, 2023a). Although alcohol or drug use decreased among offenders from 68% in 2008/09 to 46% in 2021/22, it is still a significant public health concern (The Scottish Government, 2023), with its impacts varying across different geographical settings.

Social and cultural factors significantly influence alcohol consumption patterns. These factors include religious practices (Michalak et al., 2007), familial and community ties (Bryden et al., 2013), economic circumstances (Shortt et al., 2018), ethnic background (Donath et al., 2011), general and mental health, diet and nutrition (Mohamed and Ajmal, 2015), alcohol availability (Richardson et al., 2015, Spoerri et al., 2013, Stockwell et al.,

2011), and consumption-related regulations (Bhattacharya, 2023b, Giles et al., 2024, Livingston et al., 2023). For instance, strong religious beliefs and cultural norms prohibiting alcohol consumption are closely linked to higher rates of abstinence (Michalak et al., 2007, Donath et al., 2011, Dixon and Chartier, 2016). Other protective factors, such as community social cohesion, support networks, active participation, and a supportive family environment, especially among young people, are associated with lower alcohol use (Bryden et al., 2013, Michalak et al., 2007). Moreover, prevailing social norms that support drinking behaviour, as well as contradictory messages about the harms and benefits of alcohol consumption, delay appropriate health-seeking behaviour and weaken community action (Morris et al., 2020).

Research has observed a distinction in how these factors differ between urban and rural areas. The studies by Dixon and Chartier (2016), Emslie et al. (2015) and Li et al. (2017) provide a foundation by examining how alcohol consumption varies among different demographic groups and the effectiveness of various control policies. These studies highlight that alcohol consumption is not uniform across different areas. The researchers found that factors such as age, race/ethnicity, and geographic region interact intricately with urban or rural residency, influencing patterns of alcohol use. Also, it has been found that rural areas, often characterised by tighter-knit communities, may exhibit different patterns of alcohol use compared to urban areas, where anonymity and diverse social structures prevail (Dixon and Chartier, 2016, Friesen and Kurdyak, 2020).

However, findings on the relationship between alcohol consumption patterns in rural and urban areas are contradictory. For instance, studies conducted in Australia consistently indicate that alcohol use and related harms within rural areas exceed those of metropolitan areas (Miller et al., 2010). Research in Scotland (Martin et al., 2019a, Martin et al., 2019b) and Germany (Donath et al., 2011) suggest that adolescents in rural or small-town areas have higher rates of alcohol consumption and binge drinking compared to their urban counterparts. Additionally, a study focused on driving under the influence (DUI) offenders in Nebraska (US) (Malek-Ahmadi and Degiorgio, 2015) found that rural DUI offenders have a higher risk of heavy alcohol use compared to their urban counterparts. However, Erskine et al. (2010) found that people residing in urban areas of England and Wales have higher alcohol-related mortality rates compared to those in rural areas, even after accounting for socioeconomic deprivation. Additionally, studies on the increase in alcohol consumption among the male population in China (Im et al., 2019) indicated that regular alcohol consumption was more common in urban areas (38%) than in rural areas (29%).

This apparent contradiction implies that understanding local consumption behaviours is fundamental for designing effective interventions tailored to specific community settings. For instance, specific approaches are needed to address the unique challenges of rural communities, such as limited access to healthcare, cultural norms that promote higher alcohol consumption, and social isolation (Miller et al., 2010, Still, 2024, Davis and O'Neill, 2022). Therefore, examining the context is important when implementing public policies in a specific country or setting (Friesen and Kurdyak, 2020).

Several studies aim to shed light on alcohol consumption patterns in Scotland (Torney et al., 2024), the influence of alcohol outlet density (Caryl et al., 2022, Richardson et al., 2015, Shortt et al., 2018, Shortt et al.,

2015), and the impact of minimum unit pricing (Bhattacharya, 2023b, Giles et al., 2024, Kwasnicka et al., 2021, Livingston et al., 2023, Manca et al., 2024a, Hughes et al., 2023), among other topics, such as alcohol and women in early midlife (see Emslie et al., 2015), alcohol and control policies (see Li et al., 2017) and pharmacological treatments for alcohol dependence (see Manca et al., 2024b). However, despite the growing body of research, a gap exists in understanding the rural Scottish community factors contributing to alcohol consumption patterns (SHAAP, 2020, Teckle et al., 2012).

The whisky industry, which flourished in the Highlands in the 1800s, deeply integrated alcohol into rural culture (SHAAP, 2020). This historical and economic significance of whisky has romanticised its place in Scottish identity, often downplaying the potential harms of chronic drinking and dependence (McDonald, 1994). In rural areas, alcohol is a widely accepted and prevalent drug, particularly during the winter months when indoor activities dominate and recreational options are limited (SHAAP, 2020, Still, 2024).

Different studies have addressed the relationship between alcohol consumption and rural Scotland. For example, Burns et al. (2002) explored the role and significance of alcohol in Highland communities, specifically examining the connection between alcohol consumption and mental health issues. They found that alcohol is deeply ingrained in Highland communities' culture and exacerbates issues like domestic violence, social alienation, and mental health problems influenced by gender norms and mental health stigma. Martin et al. (2019a, 2019b) investigated the impact of neighbourhoods on alcohol use among Scottish adolescents, focusing on urban and rural settings. They provided evidence that adolescents living in more remote rural areas had higher odds of having consumed alcohol compared to their urban counterparts. Kloep et al. (2001) researched young people’s drinking behaviours and their perspectives on alcohol consumption in rural communities in Scotland, Norway, and Sweden, finding that Scottish teenagers drink the most, with over 60% of Norwegian, 50% of Swedish, and about 35% of Scottish youth abstaining from alcohol. The authors explain the finding by associating it with the fact that Scottish teenagers are more engaged in commercial leisure activities, believe their parents are less concerned about their drinking and are more negative towards their teachers and school than Swedish and Norwegian teenagers living in rural areas.

There is also research carried out in rural settings about other topics, such as Daly’s (2014) doctoral thesis, which explores the experiences of rural mental health service users and providers in Scotland and Canada. Another example is Teckle et al. (2012) research investigating the relationship between rurality and health in Scotland, accounting for variations in individual and practice characteristics, focusing on hypertension, all-cause premature mortality, total hospital stays, and coronary heart disease (CHD) admissions.

Some of these studies, which consider rurality as a central aspect, are developed using a qualitative approach (Chandler and Nugent, 2016, Daly, 2014), focused on the relationships with mental health problems (Burns et al., 2002), or restricted to specific age groups (Kloep et al., 2001, Martin et al., 2019a, Martin et al., 2019b) or areas, such as the Highland communities (Burns et al., 2002), without considering other rural communities in Scotland (SHAAP, 2020, Still, 2024). However, none focus on harmful alcohol consumption in the adult population living in rural communities using a quantitative approach.

In 2022, there were 1,276 deaths solely attributed to alcohol consumption. Of these, 213 individuals, nearly 17%, came from regions categorised as 'remote small towns', 'accessible rural areas', and 'remote rural areas', based on the Scottish Government Urban Rural Classification (Geographic Information Science & Analysis Team, 2022). Since 2019, alcohol-related deaths have risen by 25% in urban areas and by 27% in remote and rural regions. Although the percentage increase between these areas is not vastly different, it is a significant concern due to the unique challenges that remote and rural communities face compared to urban ones (AFS, 2023).

Research indicates that individuals in Scotland's rural areas encounter unique drinking cultures, including traditional drinking practices, a scarcity of alcohol-free venues, and social stigma from the community, media, and healthcare professionals (AFS, 2023, SHAAP, 2020). People in small communities face heightened scrutiny due to reduced privacy, making them less likely to seek support services or engage with local recovery communities. Moreover, rural residents face significant challenges in accessing treatment and care, such as a shortage of specialised services and transportation difficulties. Furthermore, there is a lack of current information about the needs of local communities and the range of alcohol services available, which hinders effective support when they or a family member requires assistance. The Alcohol Focus Scotland, a registered national charity that works on the issue, considered that efforts to improve primary and community care for alcohol-related issues should be prioritised, focusing on the prevalence and progression of these problems in remote areas, along with the specific treatment and support needs of the population (AFS, 2023).

1.2 RESEARCH AIM

Although alcohol, especially whisky, holds significant cultural importance in Scottish society, particularly in rural areas, excessive consumption has profound negative impacts, including domestic violence, social alienation, and mental health issues. Research has highlighted the unique drinking cultures and challenges in rural communities, such as limited access to support services and heightened social stigma. However, there remains a gap in understanding the specific factors contributing to alcohol consumption patterns in rural Scottish communities despite the growing body of research on this issue.

To address this gap, this study explores and compares alcohol consumption patterns in Scottish rural and urban areas, focusing on the socio-demographic and lifestyle factors influencing this behaviour. Employing statistical methods and predictive modelling, the research seeks to understand the dynamics of alcohol consumption in rural communities.

The research objectives are to examine and describe alcohol consumption patterns in adults aged 25 to 64 in rural, urban areas and the overall population; determine if there are statistically significant differences in harmful alcohol consumption between these areas; identify and analyse the socio-demographic and lifestyle factors associated with harmful alcohol consumption among rural, urban, and general populations; and develop a predictive model to forecast problematic alcohol consumption within the rural population. The study will use data from the Scottish Health Survey (SHeS) series. Achieving these objectives, the study will contribute to developing tailored policies to reduce alcohol-related harm and promote health and well-being in Scotland, particularly in its rural regions.

2. METHODOLOGY

This research utilised a quantitative approach and employed statistics and data analysis methods to perform a secondary data analysis of a cross-sectional study: SHeS series. The first survey was conducted in 1995 and repeated in 1998, 2003, and 2008, with annual surveys being conducted since then (ScotCen Social Research, 2022a). This periodic data collection aims to estimate the prevalence of specific health conditions, assess associated risk factors, document related health behaviours, and track trends in population health and behaviours over time. The SHeS series is a crucial source of health data from private households in Scotland. Additionally, it analyses regional and demographic differences, comparing them with national statistics for Scotland and England.

2.1 DATA SOURCE

The data utilised in this study was obtained from the Scottish Health Survey (SHeS) series. Each survey series includes core questions and measurements (height, weight, and, when applicable, blood pressure, waist circumference, urine, and saliva samples) alongside annually varying modules on specific health conditions (ScotCen Social Research, 2017c, 2018b, 2019b, 2021c).

The Scottish Health Survey uses a two-stage clustered sample design, selecting intermediate geographies at the first stage and address points at the second. Each year, a sample of addresses is drawn from the Postcode Address File (PAF), comprising four types: the main sample with biological measures, the main sample without biological measures, the child boost screening sample, and the Health Board boost sample. The interviews were conducted in different households and with different individuals each year, reflecting a repeated cross-sectional study approach. Addresses are organised into interviewer assignments, and each sampled address receives an advance letter about the survey.

For the main and Health Board boost samples, all adults aged 16 and over in responding households are selected for interview. Interviewers first complete a household questionnaire with basic information about all household members, regardless of their eligibility for the survey. Individual questionnaires are then created for each eligible participant. Where possible, interviews are conducted with all eligible adults and children in the household. Data collection involved a computer-assisted telephone interview (CATI) and an online or paper self-completion questionnaire.

2.2 DATASET

This research utilised a combined dataset, incorporating data from all individuals interviewed as part of the SHeS between 2017 and 2021, the latest processed survey data available in the UK Data Service (UK Data Service, 2024) (see Section 2.3). However, the combined dataset does not include information from 2020 due to the COVID-19 pandemic. Fieldwork for SHeS 2020 was halted in March 2020, and the resulting data was classified as experimental and not comparable to previous years, leading to its exclusion from the four-year combined dataset.

In 2021, SHeS was also affected by the pandemic (ScotCen Social Research, 2021c). The interviews were conducted by telephone due to COVID-19 and social distancing measures. Fieldwork faced significant disruption, leading to two phases of data collection. In Phase 1, potential participants received invitation letters and opted for phone interviews, not in-person as usual. This phase ran from April to September 2021 and included comprehensive content like previous years. Phase 2 started in late October 2021, after COVID-19 restrictions were lifted, with approval from Scottish Government ministers and the Chief Medical Officer. During this second part, a 'knock-to-nudge' method was used where interviewers contacted potential respondents at their homes to encourage participation. Despite this doorstep contact, interviews were still conducted by phone.

Also, it is important to mention that following the 2016 consultation on the Scottish Health Survey (ScotCen Social Research, 2016) and the 2017 Scottish Government review (ScotCen Social Research, 2017a), several key changes were implemented for the 2018-2021 surveys. These include an increased sample size for Local Authority level analysis, removal of certain questions and the urine sample from the biological module, rotation of specific modules to appear biennially (including problem drinking), and the introduction of new questions on public service satisfaction, Nicotine Replacement Therapy (NRT), asthma, diabetes, and gender identity.

The consolidated dataset comprises information from household questionnaires, main individual schedules, and self-completion forms, providing a robust foundation for analysing various variables. This combined dataset comprises 25,128 records and 1,780 variables (ScotCen Social Research, 2021a).

2.3 DATA COLLECTION

The raw data file SHeS17181921.sav, the combined dataset to be processed, was downloaded from the UK Data Service (UK Data Service, 2024). The latest processed survey data was from 2022 (The Scottish Government, n.d.-a). However, only summary tables were available for 2022, and the raw data itself was not accessible despite indications in the technical report that it would be obtainable via the UK Data Service (ScotCen Social Research, 2022b). A request was submitted to the UK Data Service for the raw data from 2022. Nonetheless, the response indicated the data was unavailable (Annex A: Copy of UK Data Service's email).

The most recent raw data available on the UK Data Service was from 2021 UK (UK Data Service, 2024). This dataset was significantly affected by challenges encountered during the 2021 survey, primarily due to the COVID-19 pandemic (see Section 2.2). It was considered that modifications in the administration method and sampling procedures for this version could influence the results obtained and affect comparability with previous SHeS data (ScotCen Social Research, 2022b).

Consequently, it was decided to use the combined data from 2017-2021, which excludes data from SHeS 2020 (see Section 2.2). The UK Data Service provided the dataset (ScotCen Social Research, 2023) after submitting a required application and accepting the terms of use (UK Data Service, 2024). The combined dataset 2017-21 ensured a more consistent and reliable dataset, mitigating potential biases introduced by the unique circumstances of the 2021 survey.

2.4 PRELIMINARY DATA PROCESSING AND VARIABLE SELECTION

The initial data cleaning steps were conducted using IBM SPSS Statistics for Macintosh, Version 29.0.2.0 (IBM Corp., 2023).

As mentioned previously (see Section 1.2), the present study focuses on adults aged 25 to 64; it does not incorporate young adults aged 16 to 24 or individuals over 64 years old (NHS, 2021), as recorded in the SHeS 2017-2021 dataset. Considering this age range of interest (25 to 64, inclusive), the relevant cases were selected, resulting in a dataset with 11,185 cases. This focus on the 25-64 age group allows for a more detailed analysis within this specific demographic, which is part of the broader age range covered in the original SHeS dataset.

An initial selection of variables from the original dataset, including those containing demographic, socioeconomic, health, and alcohol consumption pattern information, was made. All the variables chosen at this stage had been identified in previous research as significant factors in harmful alcohol consumption (Shortt et al., 2015, Shortt et al., 2018, Scottish Health Action on Alcohol Problems (SHAAP), 2020, Morris et al., 2020, Mohamed and Ajmal, 2015, Miller et al., 2010, Michalak et al., 2007, McDonald, 1994, Martin et al., 2019b, Martin et al., 2019a, Malek-Ahmadi and Degiorgio, 2015, Institute of Alcohol Studies (IAS), 2020, Erskine et al., 2010, Burns et al., 2002, Bryden et al., 2013, Borders and Booth, 2007). Additionally, only variables from the individual questionnaire, not the household questionnaire, were considered. Also, variables with more than 5% missing data were dropped.

After the data were cleaned and the preliminary variables of interest were selected from the raw data, a working dataset was created (overlim_data.sav) and imported as SPSS files into R Statistical Software, Version 2024.04.0+735 (R Development Core Team, 2024). All subsequent statistical processing was performed using R Statistical Software.

2.5 OUTPUT AND VARIABLES SELECTED

The selected output variable, 'overlim15,' classifies drinking behaviour based on an estimation of the weekly alcohol consumption units reported by respondents. To estimate this value, participants aged 16 years and over were asked if they consumed alcohol (ScotCen Social Research, 2017c, 2018b, 2019b, 2021c). Those who affirmed were then queried on their frequency of consumption over the past 12 months for six types of alcoholic drinks:

- Normal beer, lager, stout, cider, and shandy
- Strong beer, lager, stout, and cider
- Sherry and Martini
- Spirits and liqueurs
- Wine
- Alcopops (alcoholic soft drinks)

Subsequent questions determined the typical quantity consumed per occasion. These amounts were then converted into units of alcohol. The SHeS series employs a standardised method (ScotCen Social Research,

2017c, 2018b, 2019b, 2021c) for converting the volume of alcohol consumption reported by participants into units of alcohol (see Annex B). This conversion accounts for the alcohol concentration of each beverage, considering that, according to the UK standard, a unit of alcohol contains 8 grams or 10 millilitres of ethanol (pure alcohol) (Department of Health England, 2016).

The average number of drinking days per week was estimated following the participants' responses. The units consumed were then multiplied by the reported frequency to estimate weekly consumption. Subsequently, the participants' alcohol consumption behaviour was classified according to the weekly limits established by the UK Chief Medical Officers' Low-Risk Drinking Guidelines (Department of Health England, 2016, Department of Heath and Social Care, 2016). These guidelines, applicable to both men and women, recommend not exceeding 14 units of alcohol per week to minimise health risks.

In the original dataset (ScotCen Social Research, 2023), the output variable 'overlim15' is categorised into two groups: 'From 0 up to and including the weekly limit' (value 0) and 'Over the weekly limit' (value 1), with the weekly limit set at 14 units. Thus, the response variable 'overlim15' is defined as a factor with two levels: 'no' for cases within the weekly limit (set as the base level) and 'yes' for cases exceeding the weekly limit. Throughout the rest of the report, the 'yes' category (exceeding the weekly limit) may also be referred to as harmful alcohol consumption.

Two new variables were created by merging variables from the original dataset: urbrur_all and cig.

The SHeS data include a variable used to classify whether participants come from urban or rural areas based on different versions of the Scottish Government urban-rural classification published by Geographic Information Science, with varying names depending on the version used:

- In SHeS 2017, the variable was labelled as Urbrur2a and was based on the Scottish Government urban-rural classification version 2013/14 (Geographic Information Science & Analysis Team, 2014).
- In SHeS 2018 to 2019, the variable was labelled as Urbrur2a_16 and was based on the Scottish Government urban-rural classification version 2016 (Geographic Information Science & Analysis Team, 2018)
- In SHeS 2021, the variable was labelled as Urbrur2a_20 and was based on the Scottish Government urban-rural classification version 2020 (Geographic Information Science & Analysis Team, 2022).

Hence, the combined dataset includes three variables (Urbrur2a, Urbrur2a_16, and Urbrur2a_20), each reflecting the rural-urban classification according to different versions of the Scottish Government urban-rural classification (Geographic Information Science & Analysis Team, 2014, 2018, 2022) Since these variables are specific to certain years, they appear as missing data for years outside their designated periods. For example, the variable 'Urbrur2a' contains data only for 2017 and shows missing cases from 2018 to 2021. However, the apparent gaps are not due to missing information but because each variable represented the same rural-urban classification data according to different versions of the classification guide. A new variable, 'urbrur_all', was

created by merging these three variables. This consolidation ensures that the information is comprehensive and resolves the issue of missing data.

Also, two variables related to smoking habits were selected from the raw data:

- Cigarette smoking status (cigst2), with four categories: non-smokers, light, moderate and heavy smokers.
- Cigarette smoking status (cigst3), with three categories: non-smokers, ex-smokers and current smokers.

These two variables were merged into a newly created variable, cig, to provide comprehensive information about smoking habits. This new variable allows us to identify if current smokers are light, moderate, or heavy smokers and to distinguish between non-smokers and ex-smokers.

Table 2.1 presents the potential predictors, with the first category mentioned in each variable set as the base level. The variables were divided into two sets: one for the nominal variables and another for the ordinal variables.

TABLE 2.1 VARIABLES PRELIMINARY SELECTED FOR THE ANALYSIS

Name	Description	Categories	Type
1. Year (syear)	Survey Year	▪ 2017 ▪ 2018 ▪ 2019 ▪ 2021	Ordinal
2. Sex (sex)	Sex of respondent	▪ Male ▪ Female	Nominal
3. Age (ag16g10)	Age 16+ in 10-year bands	▪ 25-34 ▪ 35-44 ▪ 45-54 ▪ 55-64	Ordinal
4. Birthplace (birthpla3)	Country of birth	▪ Scotland ▪ Rest of the UK ▪ Elsewhere	Nominal
5. Ethnicity (ethnic05)	Ethnic background	▪ White: Scottish ▪ White: rest of the UK ▪ White: Other ▪ Asian ▪ Other minority ethnics	Nominal
6. Religion (religi04)	Religion, religious denomination or body	▪ None ▪ Church of Scotland ▪ Roman Catholic ▪ Other Christian ▪ Another religion	Nominal

TABLE 2.1 VARIABLES PRELIMINARY SELECTED FOR THE ANALYSIS

Name	Description	Categories	Type
7. Marital status (maritalg)	Person's legal relationship status	▪ Married/civil partnership ▪ Living as married ▪ Single ▪ Separated ▪ Divorced or dissolved ▪ Widowed	Nominal
8. Life satisfaction (lifesat2)	How satisfied with life as a whole nowadays	▪ Above mode (mode = 8) ▪ Mode (mode = 8) ▪ Below mode (mode = 8)	Ordinal
9. General health (genhelf)	Self-assessed general health	▪ Very good ▪ Good ▪ Fair ▪ Bad ▪ Very bad	Ordinal
10. LTC (limitac_h)	Whether any Long-Term Condition (LTC) limits activities (DH/Long Term Conditions, 2012)	▪ No limitations ▪ Not at all ▪ A little ▪ A lot	Ordinal
11. Activity level (adt10gpW)	Physical Activity level (Department of Health and Social Care, 2011)	▪ Meets recommendations ▪ Some activity ▪ Low activity ▪ Very low activity	Ordinal
12. Smoking (cig)	Person's smoking behaviour or habits	▪ Never smoked ▪ Ex-smoker ▪ Light smokers ▪ Moderate smokers ▪ Heavy smokers	Ordinal
13. Educational level (hedqul08)	Highest educational qualification	▪ Degree or higher ▪ HNC/D ▪ Higher grade or equivalent ▪ Standard grade or equivalent ▪ Other school level ▪ No qualifications	Ordinal

TABLE 2.1 VARIABLES PRELIMINARY SELECTED FOR THE ANALYSIS

Name	Description	Categories	Type
14. Social class (schrpg7)	National Statistics Socio-economic Classification (NS-SEC) (Office of National Statistics, n.d.)	▪ I Professional ▪ II Managerial technical ▪ IIIN Skilled non-manual ▪ IIIM Skilled manual ▪ IV Semi-skilled manual ▪ V Unskilled manual ▪ Others	Nominal
15. Economy activity (neconacb)	Person's current participation in the labour market	▪ In employment ▪ ILO unemployed ▪ Inactive	Nominal
16. SIMD quintile (simd20_rpa)	Scottish Index of Multiple Deprivation (The Scottish Government, n.d.-b)	▪ Least deprived ▪ 4th ▪ 3rd ▪ 2nd ▪ Most deprived	Ordinal
17. Alcohol units (drating)	Total Units of alcohol/week	-	Continuous
18. Beers (nberwu)	Units of normal beer/week[b]	-	Continuous
19. Strong beers (sberwu)	Units of strong beer/week[b]	-	Continuous
20. Spirits (spirwu)	Units of spirits/week[b]	-	Continuous
21. Sherry (sherwu)	Units of sherry/week[b]	-	Continuous
22. Wine (winewu)	Units of wine/week[b]	-	Continuous
23. Alcopops (popswu)	Units of alcopops/week[b]	-	Continuous
24. Urban-rural class (urbrur_all)	Scottish Government urban-rural classification (Geographic Information Science & Analysis Team, 2014, 2018, 2022)	▪ Urban ▪ Rural	Nominal

[a] Annex B: Conversion used for the SHeS to convert volumes of alcohol reported in the survey into units.

The variables were converted to factors using the labels previously assigned in SPSS and numeric format, where applicable, to facilitate further analysis. The levels of factor variables were thoroughly checked. For ordinal variables, the levels' order was consistently modified when required to reflect their inherent meaning. All variable names were standardised by converting them to lowercase (Venables and Smith, 2024, Wickham,

2019). Additionally, factor levels were cleaned and standardised using the created factornames_fx function, which replaced spaces and special characters with underscores and converted them to lowercase.

Each variable within this nominal list was converted to a factor using a loop, ensuring that the data was appropriately treated as a nominal categorical variable in subsequent analyses. Secondly, the list of ordinal variables was also converted to factors within a loop; however, they were specifically designated as ordered factors in this case. The dataset's factor levels and structure were checked before modelling.

2.6 URBAN AND RURAL DATASETS

The dataset containing all the cases was filtered to select the urban and rural population groups, using the variable urbrur_all (Table 2.1) that classifies the population as urban or rural following the Scottish Government Urban Rural Classification (Geographic Information Science & Analysis Team, 2014, 2022). As a result of this step, three different datasets were obtained to perform the rest of the analysis:

- A dataset that contains cases from the urban and rural population data (dw), and since now, named urban & rural dataset.
- A dataset that contains cases from the urban population data (dwu), and since now, named urban dataset.
- A dataset that contains cases from the rural population data (dwr), and since now, named rural dataset.

2.7 HANDLING WITH HIGHLY CORRELATED VARIABLE AND LOW REPRESENTATION VARIABLE.

The variables Total Units of alcohol/week (drating), Units per week of normal beer (nberwu), strong beer (sberwu), spirits (spirwu), sherry (sherwu), wine (winewu), and alcopops (popswu) were excluded for further analysis due to their high correlation with the outcome variable (overlim15). These variables quantify the total alcohol consumption per week, while overlim15 is a binary classification indicating whether weekly alcohol consumption exceeds the recommended 14-unit limit based on frequency (see Section 2.5). Given the redundancy between these variables and the outcome, including them in the model would introduce multicollinearity, which could distort the model's parameter estimates. Therefore, these variables were deemed unnecessary for the analysis.

Upon examining the original dataset, it was found that some categories within certain variables had very low representation in the urban & rural, urban, and rural datasets. To address this, recoding was performed to enhance the representation of these categories (Table 2.2). Variables affected included ethnicity (rthnic05), religion (religi04), marital status (maritalg), educational level (hedqul08), social class (schrpg7), individual economic activity (neconacb), and activity level (adt10gpW). These adjustments were essential to facilitate robust analysis and ensure the statistical significance of the findings.

TABLE 2.2 RE-CODING OF UNDER-REPRESENTED CATEGORIES

Before re-coding[a]				After re-coding[a]			
Variables and categories	Datasets			Categories	Datasets		
	Urban & Rural	Urban	Rural		Urban & Rural	Urban	Rural
Ethnicity (rthnic05) – N[b] = 11,150							
White-Scottish	8,394 (75%)	6,804 (76%)	1,590 (72%)	White-Scottish	8,394 (75%)	6,804 (76%)	1,590 (72%)
White-Rest UK	1,568 (14%)	1,070 (12%)	498 (22%)	White-Rest UK	1,568 (14%)	1,070 (12%)	498 (22%)
White-Other	690 (6.2%)	586 (6.6%)	104 (4.7%)	Other	1,188 (11%)	1,059 (12%)	129 (5.8%)
Asian	304 (2.7%)	291 (3.3%)	13 (0.6%)				
Other	194 (1.7%)	182 (2.0%)	12 (0.5%)				
Religion (religi04) – N[b] = 11,145							
None	6,127 (55%)	4,853 (54%)	1,274 (58%)	none	6,127 (55%)	4,853 (54%)	1,274 (58%)
Church of Scotland	2,262 (20%)	1,751 (20%)	511 (23%)	Church of Scotland	2,262 (20%)	1,751 (20%)	511 (23%)
Roman Catholic	1,508 (14%)	1,329 (15%)	179 (8.1%)	Roman Catholic	1,508 (14%)	1,329 (15%)	179 (8.1%)
Other Christian	893 (8.0%)	667 (7.5%)	226 (10%)	Other religion	1,248 (11%)	998 (11%)	250 (11%)
Another religion	355 (3.2%)	331 (3.7%)	24 (1.1%)				
Marital status (maritalg) – N[b] = 11,183							
Married-Partner	6,045 (54%)	4,656 (52%)	1,389 (63%)	Married-Partner	6,045 (54%)	4,656 (52%)	1,389 (63%)
As married	1,705 (15%)	1,375 (15%)	330 (15%)	As married	1,705 (15%)	1,375 (15%)	330 (15%)
Single	2,046 (18%)	1,768 (20%)	278 (13%)	Single	2,046 (18%)	1,768 (20%)	278 (13%)
Separated	346 (3.1%)	292 (3.3%)	54 (2.4%)	Separated-Widowed	1,387 (12%)	1,164 (13%)	223 (10%)
Divorced-Dissolved	842 (7.5%)	714 (8.0%)	128 (5.8%)				
Widowed	199 (1.8%)	158 (1.8%)	41 (1.8%)				
General health (genhelf2) – N[b] = 11,180							
Very good	3,928 (35%)	3,089 (34%)	839 (38%)	Very good	3,928 (35%)	3,089 (34%)	839 (38%)
Good	4,397 (39%)	3,506 (39%)	891 (40%)	Good	4,397 (39%)	3,506 (39%)	891 (40%)
Fair	1,930 (17%)	1,582 (18%)	348 (16%)	Fair	1,930 (17%)	1,582 (18%)	348 (16%)
Bad	693 (6.2%)	586 (6.5%)	107 (4.8%)	Bad-Very bad	925 (8.3%)	784 (8.7%)	141 (6.4%)
Very bad	232 (2.1%)	198 (2.2%)	34 (1.5%)				

TABLE 2.2 RE-CODING OF UNDER-REPRESENTED CATEGORIES

Before re-coding[a]				After re-coding[a]			
Variables and categories	Datasets			Categories	Datasets		
	Urban & Rural	Urban	Rural		Urban & Rural	Urban	Rural
Educational level (hedqul08) – N[b] = 11,137							
Degree or Higher	4,812 (43%)	3,814 (43%)	998 (45%)	Degree or Higher	4,812 (43%)	3,814 (43%)	998 (45%)
HNC/D	1,646 (15%)	1,327 (15%)	319 (14%)	HNC/D	1,646 (15%)	1,327 (15%)	319 (14%)
Higher grade	1,580 (14%)	1,232 (14%)	348 (16%)	Higher grade	1,580 (14%)	1,232 (14%)	348 (16%)
Standard grade	1,890 (17%)	1,553 (17%)	337 (15%)	Standard-school grade	2,078 (19%)	1,720 (19%)	358 (16%)
Other school level	188 (1.7%)	167 (1.9%)	21 (0.9%)				
No qualifications	1,021 (9.2%)	830 (9.3%)	191 (8.6%)	No qualifications	1,021 (9.2%)	830 (9.3%)	191 (8.6%)
Social class (schrpg7) – N[b] = 11,029							
Professional	1,091 (9.9%)	880 (10.0%)	211 (9.6%)	Professional	1,091 (9.9%)	880 (10.0%)	211 (9.6%)
Managerial technical	4,380 (40%)	3,429 (39%)	951 (43%)	Managerial technical	4,380 (40%)	3,429 (39%)	951 (43%)
Skilled non-manual	1,517 (14%)	1,280 (14%)	237 (11%)	Skilled non-manual	1,517 (14%)	1,280 (14%)	237 (11%)
Skilled manual	1,896 (17%)	1,506 (17%)	390 (18%)	Skilled manual	1,896 (17%)	1,506 (17%)	390 (18%)
Semiskilled manual	1,463 (13%)	1,173 (13%)	290 (13%)	Semiskilled manual	1,463 (13%)	1,173 (13%)	290 (13%)
Unskilled manual	467 (4.2%)	375 (4.2%)	92 (4.2%)	Unskilled-other	682 (6.2%)	565 (6.4%)	117 (5.3%)
Others	215 (1.9%)	190 (2.2%)	25 (1.1%)				
Economic activity (neconacb) – N[b] = 11,165							
In employment	8,446 (76%)	6,698 (75%)	1,748 (79%)	In employment	8,446 (76%)	6,698 (75%)	1,748 (79%)
ILO unemployed	267 (2.4%)	228 (2.5%)	39 (1.8%)	Unemployed-inactive	2,719 (24%)	2,249 (25%)	470 (21%)
Inactive	2,452 (22%)	2,021 (23%)	431 (19%)				
Activity level (adt10gpW) – N[b] = 11,121							
Meets recommendations	7,870 (71%)	6,231 (70%)	1,639 (74%)	Meets recommendations	7,870 (71%)	6,231 (70%)	1,639 (74%)
Some activity	1,103 (9.9%)	903 (10%)	200 (9.1%)	Low activity	1,497 (13%)	1,221 (14%)	276 (12%)
Low activity	394 (3.5%)	318 (3.6%)	76 (3.4%)				
Very low activity	1,754 (16%)	1,460 (16%)	294 (13%)	Very low activity	1,754 (16%)	1,460 (16%)	294 (13%)

[a] The percentages are calculated with respect to the total of each category in each dataset.

[b] N of the urban & rural dataset.

2.8 SPLITTING THE DATASET IN TRAINING AND TEST SUBSETS.

The three datasets, with urban & rural cases (dw), with urban cases (dwu), and with rural cases (dwr), were divided into training and test subsets for proper statistical model training, evaluation, and validation (Peng et al., 2002, Louppe, 2014). A random seed (i.e. 123) was set to ensure the reproducibility of the results. The createDataPartition function from the caret package (Kuhn, 2008) was utilised to ensure that the distribution of the dependent variable ('overlim15') was preserved across both subsets, allocating 70% of the data to the training subset and the remaining 30% to the test subset (see Section 3.3).

All processing steps performed on the training subset were replicated on the test subset to maintain integrity, allowing for accurate and reliable comparisons between model performance on the training and test data.

2.9 HANDLING THE MISSING DATA

The original SPSS dataset contained missing values labelled as follows:

- -1: Not applicable.
- -2: Schedule not applicable.
- -6: Schedule not obtained.
- -8: Don't know.
- -9: Refused.
- -90: Age of household member refused
- -99: Unclassifiable

All these missing values were converted to 'Not Available' (NA) and considered missing data.

The proportion of missing data for each variable in the working datasets was calculated using the sapply function (Wickham, 2019). This function is designed to work with lists or vectors and apply a specified function to each element of the list or vector, simplifying the output. Additionally, visual representations of the missing data were generated.

The method employed for handling missing data was listwise deletion (also known as complete case analysis) (Allison, 2002). This technique excludes any rows (cases) from the dataset containing missing values, refining the dataset to include only observations with complete data across all variables. This procedure was applied after dividing the data into training and test subsets (see Section 2.8), and it was performed on both groups of subsets separately (training and test subsets), ensuring they contained only complete cases (see Section 3.4).

Accomplishment of listwise deletion after the split ensures that the training and testing datasets maintain their intended proportions and characteristics, preventing any bias that might arise from removing cases before the split. This procedure enhances the robustness and validity of subsequent statistical analyses by ensuring that each subset accurately represents the overall dataset's structure.

2.10 HANDLING THE CLASS IMBALANCE

As shown in Table 2.3, the output variable overlim15 was imbalanced in the three training subsets: dw.train, dwu.train, and dwr.train. In each of these datasets, the majority class (No) significantly outnumbered the minority class (Yes), with approximately a 3:1 ratio.

TABLE 2.3 CLASS IMBALANCE IN THE OUTPUT VARIABLE (overlim15)

Datasets	Cases	Output Variable: overlim15[a]	
		Majority class (No)	Minority class (Yes)
dw.train	7,548	5,741 (76.06%)	1,807 (23.94%)
dwu.train	6,036	4,587 (75.99%)	1,449 (24.01%)
dwr.train	1,498	1,138 (75.97%)	360 (24.03%)

[a] The percentages are calculated concerning the number of cases in each dataset.

This imbalance in the output variable across all training subsets was considered to affect the robustness and reliability of the predictive models developed from these subsets (Japkowicz, 2000, López, 2013). To address this imbalance, oversampling (see Section 2.10.1) and the Synthetic Minority Over-sampling Technique (see Section 2.10.2) were employed. These techniques aim to balance the class distribution by either increasing the number of minority class instances or generating synthetic samples respectively.

2.10.1 OVERSAMPLING

An oversampling procedure was employed to address the class imbalance within the output variable overlim15. The primary aim was to create balanced datasets by increasing the representation of the minority class (Yes) into three specific proportions relative to the majority class (No) (Japkowicz, 2000, López, 2013):

- 30% minority and 70% majority, ratio 0.3.
- 40% minority and 60% majority, ratio 0.4.
- 50% minority and 50% majority, ratio 0.5.

To facilitate this process, an oversampling function was developed. The function first set a random seed for reproducibility and defined the desired proportional ratios (0.3, 0.4, and 0.5) for the minority class. For each dataset, the function splits the data into majority and minority subsets based on the target variable overlim15. For each specified ratio, the necessary count of minority class samples was calculated (1) to achieve the desired proportion:

$$\text{Target minority count} = \frac{\text{proportion ratio} \times \text{nrow(majority class)}}{\text{1- proportion ratio}} \quad (1)$$

The minority class samples were randomly selected, seting a random seed, with replacements to reach the target count, ensuring a sufficiently large minority class subset. The original majority class data was combined with the unsampled minority data to form a balanced dataset for each specified ratio. This oversampling function was applied to the training sets (dw.train, dwu.train and dwr.train) of the urban & rural, urban, and rural datasets using lapply, a function in R that applies a specified function over a list and returns a list of results (Wickham, 2019). Each resultant dataset (see Section 3.5) was named according to its respective oversampling

ratio and the cases subgroup (urban & rural, urban, or rural) to facilitate easy identification and subsequent analysis.

2.10.2 SMOTE

Also, to address the class imbalance in the training datasets (dw.train, dwu.train and dwr.train), it was employed the Synthetic Minority Oversampling Technique (SMOTE) (Elreedy and Atiya, 2019, Fernández, 2018, Chawla Nitesh, 2002, López, 2013). SMOTE is an oversampling technique that creates synthetic samples for the minority class to balance the class distribution, which is crucial for improving the predictive model's performance and reliability. Each minority sample's k-nearest neighbours are identified, and synthetic examples are created by interpolating between the feature values of the original sample and its selected neighbours.

A custom SMOTE function was developed to automate this process. At the beginning of the processing, a seed for reproducibility was set. The k parameter in SMOTE determines how many nearest neighbours are used to generate the synthetic samples, with higher values potentially producing more diverse samples. For this study, three k values (3, 5, and 10) were tested, representing the nearest neighbours considered during synthetic sample generation (Chawla Nitesh, 2002). Also, during the SMOTE, the over-sampling ratio (or) should be established because this specifies the proportion of the synthetic samples generated compared to the majority class. In this case, three different or were defined: 0.5, 0.75, and 1.

A dataset was created for each combination of k and or. The SMOTE process was applied separately to the training subsets (dw.train, dwu.train and dwr.train) for the urban & rural, urban, and rural population groups, resulting in 27 different datasets. Each resulting dataset from the SMOTE application was given a unique name reflecting the k value and over-sampling ratio used (see Section 3.6). These datasets were stored in a list for easy access and further analysis.

2.11 METHODS

This section outlines the methods employed in the study. Specifically, it describes using descriptive statistics and visualisations, chi-squared test, logistic regression analysis, and random forest classification. These methods examined the relationships between various demographic and socioeconomic factors and alcohol consumption behaviour in the three datasets: the urban & rural, urban, and rural population groups.

2.11.1 DESCRIPTIVE STATISTICS AND VISUALISATIONS

Descriptive statistics and visualisations were conducted to comprehensively understand alcohol consumption patterns in the urban & rural, urban and rural population groups. These methods facilitated the exploration of the distribution of alcohol consumption data, identifying potential trends and differences (Wilkinson, 2005, Wickham et al., 2023, Wickham, 2019).

The proportion of each categorical variable in the urban & rural, urban, and rural population groups was obtained. Means ($\bar{x}$), medians (Med), and standard deviations (SD) were computed for the numeric variables to summarise the central tendency and dispersion of alcohol consumption within the urban & rural, urban and

rural population groups. The same statistical summaries were obtained by grouping the data by the categorical outcome variable (overlim15).

To visualise the data, a custom function, 'fx_plot.cat', was defined to create bar plots for the categorical variables of interest:

- Sociodemographic variables (e.g., age, sex, country of birth),
- Economic situation variables (e.g., economic activity, SIMD quintiles),
- Health variables (e.g., self-assessed health, life satisfaction), and
- Habit variables (e.g., smoking and drinking behaviours).

These plots compared distributions considering the urban/rural classification. For numerical variables related to alcohol consumption, box plots were created. These plots utilised log-transformed values to enhance the visualisation and interpretation of the data distributions.

2.11.2 CHI-SQUARED TEST

A chi-squared test, with a significance level of $\alpha = 0.05$, was conducted to determine if there was a significant association between the binary categorical variables 'overlim15' (categorising alcohol consumption) and 'urbrur_all' (urban or rural classification). The observed and expected frequencies and residuals were calculated. Additionally, a correlation plot was used to represent the residuals visually.

The chi-squared test is commonly used to determine whether there is a significant association between two categorical variables by comparing the observed frequencies in each category with the expected frequencies, which would occur if there were no associations between the variables (Nikulin and Chimitova, 2017).

The test statistic for the chi-squared test was calculated as follows:

$$X^2 = \sum \frac{(O_i - E_i)}{E_i}$$

where O_i represents the observed frequency in each category, and E_i represents the expected frequency under the null hypothesis of no association.

The resulting chi-squared statistic was then compared to a critical value from the chi-squared distribution with the appropriate degrees of freedom (df), which is calculated as:

$$df = (r - 1)\,(c - 1)$$

where r is the number of rows and c is the number of columns in the contingency table.

This study used the chi-squared distribution to determine the statistical significance of the association between urbrur_all and overlim15 variables in unpaired data, considering the following hypotheses:

- Ho: No association exists between exceeding the weekly alcohol consumption limit and living in a rural or urban community.

- Ha: There is an association between exceeding the weekly alcohol consumption limit and living in a rural or urban community.

A p-value less than the predetermined significance level ($\alpha = 0.05$) was used to reject the null hypothesis of no association between the categorical variables.

The following assumptions (McHugh, 2013) were checked and fulfilled:

- Frequency data: The data in each cell were counts of cases, not percentages or other transformations.
- Categorical variables with mutually exclusive categories: The variables urbrur_all and overlim15 were categorical with two mutually exclusive categories, meaning each case fit into only one category for each variable.
- Independent contributions: Each case contributed data to only one cell in the Chi-squared table; there were no repeated measures on the same case over time.
- Cell expected values: At least 80% of the cells had expected frequencies of five or more, and no cell had an expected frequency of less than one.

2.11.3 LOGISTIC REGRESSION

To examine the relationship between harmful alcohol consumption and various demographic and socioeconomic factors among individuals in rural and urban settings, a multivariate logistic regression (LR) analysis was used (Ledolter, 2013, Peng et al., 2002). This method is well-suited for modelling binary categorical outcome (dependent) variables using multiple predictors. In this case, the dependent variable (overlim15) classified individuals based on whether they exceeded or did not exceed the weekly alcohol consumption limit (Department of Health England, 2016, Department of Heath and Social Care, 2016).

LR is widely applied in fields such as finance, healthcare, and social sciences to predict the likelihood of events. The objective is to model the probability that Y=1 (i.e. exceeding the limit), given a set of predictor variables X = [X_1, X_2,..., X_k], with p representing the probability of success (Y=1) (2).

$$p = P(Y = 1|X) \quad (2)$$

The LR of Y on X estimates parameter values for $\beta = \beta_0, \beta_1 \ldots \beta_k$, via the maximum likelihood method. It posits that the logit of the probability p is a linear function of the predictor variables, as described by the following equation:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k X_k$$

where β_0 is the intercept, and β_1, β_2, ..., β_k are the coefficients associated with the predictor variables X_1, X_2, ..., X_k.

In LR, the coefficients β have a specific interpretation regarding odds ratios (OR). The OR for a predictor Xj is given by [eβj], representing the change in the outcome odds for a one-unit increase in Xj. In the case of categorical variables, as in this study, the OR are interpreted relative to the base (reference) category. An OR > 1

for a particular category indicates higher odds of the outcome occurring than the base category, while an OR < 1 indicates lower odds than the base category. An OR = 1 indicates no difference in odds between the category and the base category.

Regarding the confidence intervals (CI) of the OR, if they include 1, the effect may not be statistically significant. An initial analysis was performed using all predictors (i.e., a full LR model) to examine the relationship between the outcome variable and all other independent variables (see Appendix: Table I.1) in the training datasets.

A function, fx_steplg, was created to perform stepwise logistic regression (LR) on 39 training datasets. These datasets include the original urban & rural, urban, and rural datasets (dw.train, dwu.train, and dwr.train), nine oversampled versions, and 27 generated using SMOTE, as described in Sections 2.10.1, and 2.10.2.

Stepwise selection, commonly used in statistical modelling and medical research, was chosen for its ability to iteratively refine models by adding and removing variables based on statistical significance(Chowdhury and Turin, 2020, Yamashita et al., 2007). This approach balances model complexity and fit, making it well-suited for datasets with numerous predictors, such as those in this study. The fx_steplg function employs the stepAIC method. Using the Akaike Information Criterion (AIC) (Akaike, 1974, Yamashita et al., 2007), the function refines the model by iteratively adding and removing predictors in both directions, optimising the trade-off between goodness of fit and parsimony and thereby minimising information loss.

The AIC is a measure used in statistical model selection to assess the quality of different models. Based on the concept of information entropy, it provides a relative estimate of the information lost when a given model is used to represent the process that generates the data. The AIC is calculated as:

$$\mathrm{AIC} = 2k - 2\ln(\mathcal{L})$$

where k is the number of parameters in the model, and $\mathcal{L}$ is the maximum likelihood of the model, i.e., the highest probability that the model can explain the observed data obtained by maximising the likelihood function (3) or, in practice, the natural logarithm of the likelihood function (4),

$$L(\theta) = P(X|\theta) \tag{3}$$

$$\ln(L(\theta)) = \sum_{i=1}^{n} \ln P(X_i|\theta) \tag{4}$$

where i goes from 1 to n, n is the number of data points, and $P(X_i|\theta)$ is the probability of the observed data X_i given the parameters θ. A lower AIC value indicates a better-fitting model, considering the trade-off between goodness of fit and model complexity. The AIC is particularly useful for comparing multiple models and selecting the one that balances best fit and complexity.

The resulting coefficients were then examined, and the final selected predictors, along with their statistical significance in predicting alcohol consumption, were summarised for each of the datasets. The selected predictors were used to fit the LR model on each of the 39 training sets, ensuring that only the significant variables were included in the final models.

The LR assumptions were checked (Stoltzfus, 2011):

- First, a binary LR model was utilised, appropriate for a binary dependent variable. In this case, the variable 'overlim15' had two categories: below and above the weekly alcohol consumption limit (Department of Health England, 2016, Department of Heath and Social Care, 2016).
- Second, in the LR, the observations must be independent, meaning that the data should not come from repeated measurements or matched data. This requirement was met as the data source, the SHeS, is a cross-sectional study (see Section 2.1).
- Third, LR requires low or no multicollinearity among the independent variables, ensuring the predictors are not excessively correlated. Multicollinearity occurs when independent variables in a regression model are highly correlated, which can affect the stability and interpretability of the model coefficients. This was verified by calculating the Variance Inflation Factor (VIF) (Chatterjee and Hadi, 2006). The VIF is a measure used to detect multicollinearity in regression analysis, quantifying how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors (5).

$$\text{VIF} = \frac{1}{1 - R_i^2} \tag{5}$$

  where R_i^2 is the coefficient of determination of the regression model that predicts the ith predictor using all other predictors. A VIF value of 1 indicates no correlation, values between 1 and 5 suggest moderate correlation, and values above 5 indicate high correlation, which may require corrective measures such as removing or combining variables.
- Fourth, each model fulfilled the adequacy of sample size, often guided by the "events per variable" (EPV) rule, which suggests a minimum of 10 events per predictor variable to improve the model's convergence, fit, and generalisability. A large sample size ensures reliable and stable parameter estimation, increasing statistical power and precision. Larger samples provide the necessary occurrences of both outcomes, particularly for rare events, and mitigate issues such as multicollinearity. Additionally, they enhance the model's robustness against missing data and outliers.

To validate the LR model, a 10-fold cross-validation technique was employed using the `caret` package in R (Kuhn, 2008). This method systematically tests the model on different subsets of the data to evaluate its generalizability and robustness. Specifically, each dataset was divided into ten equally sized folds (k = 10); each iteration uses one-fold as the test set, while the remaining nine folds are used for training. This process is repeated ten times, ensuring each fold is used exactly once as the test set. This approach helps to ensure that the model performs well on unseen data and reduces the risk of overfitting. The primary metrics for evaluating model performance were:

- Receiver Operating Characteristic (ROC) curve: It is a graphical representation used to evaluate the performance of a binary classification system. It plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings.

- Area under the ROC curve (ROC AUC): It provides a single measure of overall accuracy, where a ROC AUC of 1 indicates perfect classification, and a ROC AUC of 0.5 suggests no discriminative power.
- Confusion matrix: Based on the optimal threshold indicated by the ROC curve, the confusion matrix classifies predicted probabilities into binary outcomes. This 2x2 table provided a detailed performance breakdown of a classification model by summarising the counts of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).
- Sensitivity (Recall): The proportion of TP correctly identified by the model (6).

$$\text{Sensitivity (Recall)} = \frac{TP}{FN + TP} \tag{6}$$

- Specificity: The proportion of TN correctly identified by the model (7).

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{7}$$

- Accuracy: The overall proportion of correctly classified instances (8).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

- Positive Predictive Value (PPV): The proportion of correct positive predictions (9).

$$\text{PPV} = \frac{TP}{TP + FP} \tag{9}$$

- F1-score: The harmonic means of precision and sensitivity (10).

$$\text{F1-score} = 2\left(\frac{\text{PPV} \times \text{Sensitivity}}{\text{PPV} + \text{Sensitivity}}\right) \tag{10}$$

After conducting cross-validation on the 39 models, the categories that were not significant ($p > 0.05$) were re-coded. Additionally, variables that did not contribute meaningfully to the models ($p > 0.05$) were excluded. The best-performing model for the urban & rural, urban, and rural training subsets (dw.train, dwu.train and dwr.train) was selected based on the cross-validation results, resulting in one model for each training dataset.

The selected models were then fitted using the respective training data after re-coding some of the categories as mentioned in the previous paragraph. The models were interpreted to understand the significance and impact of the predictors. Subsequently, the fitted models were evaluated using the respective test subsets (see Section 2.8) to provide an unbiased assessment of their performance. Metrics such as recall (6), accuracy (8), ROC AUC, PPV (9), and F1-score (10), initially calculated during the preliminary evaluation of the training subsets, were similarly obtained for the test subsets.

2.11.4 RANDOM FOREST FOR CLASSIFICATION

A Random Forest (RF) classification method was applied as an alternative to LR modelling (Couronne et al., 2018, Kirasich and Sadler, 2018, Muchlinski et al., 2016) on the original rural dataset (dwr), as well as on the datasets with rural cases obtained through oversampling (see Section 2.10.1) and SMOTE (see Section 2.10.2). All these rural datasets were compiled into a list, which included:

- The original rural training dataset: dwr.train
- The three upsampled rural training datasets: up30_dwr, up40_dwr and up50_dwr
- The nine SMOTE-generated rural training datasets: smk3r50_dwr, smk5r50_dwr, smk10r50_dwr, smk3r75_dwr, smk5r75_dwr, smk10r75_dwr, smk3r100_dwr, smk5r100_dwr and smk10r100_dwr

A function was then employed to separate the features (independent variables) from the output variable overlim15 in each dataset. This function iterated over each dataset, selecting all columns except the output variable (overlim15). It built two sets for each dataset: the feature set (X_train) and the output variable set (y_train), with the output variable overlim15, previously isolated. The resulting feature and output sets were stored in a list for further RF processing.

RF is a nonparametric learning method. Its core concept is to combine numerous independent decision trees, each serving as a basic classifier (Biau and Scornet, 2016, Genuer and Poggi, 2020). Due to the random variations in constructing individual trees, the forest explores a broader range of potential tree predictors, typically resulting in enhanced predictive performance. Each tree has different types of nodes (Rigatti, 2017):

- Root Node: The topmost node represents the entire dataset, which splits into two or more homogeneous sets.
- Internal Nodes: Nodes representing decisions based on specific feature values, splitting into further nodes.
- Terminal Nodes (or Leaf Nodes): The endpoints of the tree, representing the final decision or output, with no further splits. The minimum size of terminal nodes is an important parameter that helps control the complexity of the tree. Setting a minimum size for terminal nodes ensures that each leaf node contains at least a certain number of data points, which helps prevent the tree from becoming too detailed and overfitting the training data.

Considering the working dataset as a learning sample $\mathcal{L}_n$ (11) composed of n pairs of independent and identically distributed (i.i.d.) observations drawn from the same unknown distribution as the pair (X, Y), the aim is to estimate the relationship between the input variables (X) and the output or dependent variable (Y) (Genuer and Poggi, 2020).

$$\mathcal{L}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \quad (11)$$

It is assumed that $X \in \mathcal{X}$, with a dimension p, where p is the total number of input variables (or predictors); and for a binary classification problem $Y \in \mathcal{Y}$, where $\mathcal{Y}$= {1, 2}. The final goal is to build a predictor, which links a predicted value $\hat{y}$ of the output variable for any given input observation $x \in \mathcal{X}$.

$$\hat{h} = \mathcal{X} \rightarrow \mathcal{Y} \quad (12)$$

The predictor is evaluated based on the prediction error (also known as generalisation error), which is related to the probability of misclassification in the context of a classification problem and can be expressed as:

$$P\left(Y \neq \hat{h}(X)\right)$$

where $\hat{h}(X)$ is the predicted class for the input observation X. Since this depends on the unknown joint distribution of the random pair (X, Y), it needs to be estimated.

The final class label for the sample is determined by aggregating the classifications from all the individual trees through a majority voting process (Biau and Scornet, 2016, Genuer and Poggi, 2020, Louppe, 2015). During the construction of each tree, approximately one-third of the training data is not used for building that tree. These unused data points are called out-of-bag (OOB) examples. After constructing each tree, the OOB data is passed through the tree to get an unbiased estimate of the classification error.

Considering a scenario with p input variables, after constructing each tree, RF for classification randomly permutes the values of the jth variable in the OOB examples. The OOB data is then passed through the corresponding tree again, and the classification result for each OOB example x_n is recorded. This process is repeated for each variable (j = 1, 2,..., p). After completing these iterations, the plurality of the OOB class votes for x_n, with the jth variable permuted, is compared to the true class label of x_n. This process can be illustrated using the scheme from Genuer and Poggi (2020), as shown in Figure 2.1.

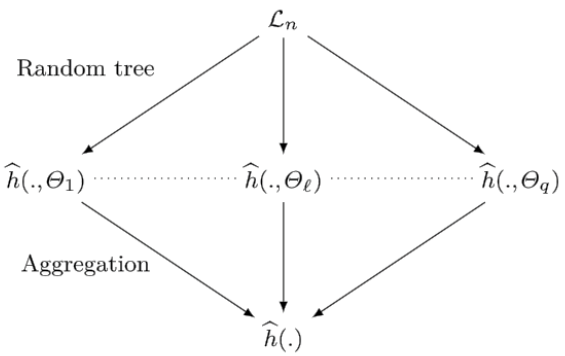


FIGURE 2.1 GENERAL SCHEME OF RF

Source: Genuer, R. and Poggi, J. (2020), Random Forests for Big Data, Big Data Journal, 8(3), pp. 123-135.

where $\left(\hat{h}(., \Theta_1),...\hat{h}(., \Theta_q)\right)$ is a collection of random trees, and $\Theta_1, \ldots \Theta_q$ are i.i.d. random variables, independent of $\mathcal{L}_n$, and used in constructing each of the q trees in the RF. The final predictor $\hat{h}(.)$ is obtained by aggregating these tree predictors.

The RF for classification method was applied to the 13 datasets containing rural cases because it is well-suited for binary classification tasks involving categorical variables (Biau and Scornet, 2016, Kirasich and Sadler, 2018, Muchlinski et al., 2016). RF handles categorical data without requiring extensive pre-processing. As with decision trees, which are the building blocks of RF, the algorithm naturally splits data based on input variable categories. By aggregating the predictions of multiple trees, RF mitigates the risk of overfitting, which is a common concern with categorical data. Being non-parametric, RF does not assume a specific data distribution, making it flexible for various data types. Additionally, it provides insights into variable importance, helping to

understand the influence of categorical predictors on binary outcomes. The ability of RF to model complex interactions between input variables further enhances its performance in binary classification tasks.

For the application of RF, the R package `ranger` was used. This package was chosen due to its superior efficiency in terms of both runtime and memory usage and its robust handling of high-dimensional datasets (Wright and Ziegler, 2017, Genuer and Poggi, 2020). The package's ability to optimise hyperparameters further enhanced its suitability for the study, which aimed to test a list of datasets, providing a reliable tool for building and evaluating RF models.

The model was trained on datasets with varying numbers of trees, numbers of variables tried at each split (mtry), minimum sizes of terminal nodes, and proportions of samples used for training sample sizes to ensure comprehensive exploration of potential hyperparameter combinations (Probst et al., 2019). To test the different combinations, a grid was constructed with the following hyperparameter values (Wright and Ziegler, 2017):

- Number of trees: 100, 200, 300 and 500. This range balances computational efficiency and model performance, ensuring sufficient trees to capture data patterns without excessive computational costs.
- mtry: 2 to 16. This range allowed assessing the full range of features considered, from a small 2 to the full model that contains 16 features, ensuring that we test models that use a small subset of features and others that consider all features at each split. Covering the full range of feature combinations ensures thorough exploration, helping to reduce overfitting and improve model robustness.
- Minimum size of terminal nodes: 3, 5, 7 and 9. Larger terminal node sizes help control overfitting, which is particularly important for small and imbalanced datasets, as in this case. A range from 3 to 9 ensures that the model can capture sufficient detail without becoming too granular and overfitting to the training data.
- Proportion of samples used for training: 0.7, 0.8, and 1. Varying sample sizes simulate bootstrap sampling effects and help reduce variance, improving model robustness on imbalanced data.

Each combination was evaluated using the OOB Root Mean Squared Error (OOB RMSE) to assess the model's prediction error. OOB RMSE is a metric used to evaluate the performance of a RF model without the need for a separate validation set (Louppe, 2015, Genuer and Poggi, 2020, Wright and Ziegler, 2017). It utilises the OOB samples (the data points not included in the bootstrap sample for a given tree) to calculate the prediction error. The OOB RMSE can be computed using the following equation:

$$\text{OOB RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i^{OOB}\right)^2}$$

where N is the total number of observations in a dataset, y_i is the actual value of the ith observation, and $\hat{y}_i^{OOB}$ is the predicted value for the ith observation using only the trees for which the ith observation was OOB. The combination yielding the lowest OOB RMSE was identified as the best set of hyperparameters for the model.

Using the `caret` package in R (Kuhn, 2008) and considering the best set of hyperparameters for the model, 10-fold cross-validation repeated five times was employed to ensure robust model evaluation. This approach helps provide a reliable estimate of model performance and mitigate overfitting. A final model was selected, and its performance was evaluated in the test subsets, calculating the recall (6), specificity (7), accuracy (8), PPV (9), and F1-score metrics (10).

3. RESULTS

This section presents the results of analyses conducted on 39 datasets derived from various oversampling techniques and SMOTE variants: 13 from the original urban & rural dataset (dw), 13 from the urban dataset (dwu), and 13 from the rural dataset (dwr).

Firstly, socio-demographic, health-related, educational, and socioeconomic variables were analysed across these three original datasets (see Section 3.1). The analysis covered gender (sex), age groups (ag16g10), birthplace (birthpla3), ethnicity (ethnic05), religion (religi04), marital status (maritalg), health status (genhelf), long-term conditions (limitac_h), life satisfaction (lifesat2), activity levels (adt10gptw), smoking habits (smoking), educational levels (hedqul08), employment status (neconacb), social class classification (schrpg7), and Scottish Index of Multiple Deprivation-SIMD (simd20_rpa). These variables were examined in relation to cases without any filter (dw) and cases filtered by urban-rural condition (dwu and dwr). Additionally, alcohol consumption behaviours about the recommended weekly limit of 14 units (Department of Health England, 2016, Department of Heath and Social Care, 2016) were explored to understand how these factors interacted across different geographic contexts (see Section 3.2). Detailed counts (n) and proportions for all categories were presented in Appendix Table I.1 for the urban & rural (dw), urban (dwu), and rural (dwr) datasets, while Appendix Table I.2 provided summaries for alcohol consumption below and above the recommended limit (overlim15).

Secondly, missing data was analysed to assess patterns and proportions of missing values across the three datasets. This step ensured the robustness of subsequent analyses and provided insights into potential biases or areas for data imputation. Thirdly, the chi-squared test results were presented, exploring the association between weekly alcohol consumption and residence in rural or urban areas (see Section 3.7). Fourthly, the predictive models' results were detailed, with separate subsections for each population group under study. Results were first provided for the 13 datasets containing urban & rural cases (see Section 3.8), then for the 13 datasets containing urban cases only (Section 3.9), and finally for the 13 datasets containing rural cases only (see Section 3.10). This section concluded with results from the random forest model (see Section 3.12), developed solely with datasets representing rural populations.

3.1 CHARACTERISTICS OF THE URBAN & RURAL, URBAN AND RURAL DATASETS

The following plots illustrate the distribution of socio-demographic, health-related, and economic variables across three datasets: urban & rural, urban only, and rural only. A detailed breakdown of quantities and percentages is provided in Appendix Table I.1.

The youngest age group (25 to 34 years) is the least represented, and had a higher presence in the urban & rural (19.52%) and urban (20.89%) datasets compared to the rural dataset (13.96%) (Figure 3.1). Conversely, the oldest age group (55 to 64 years) is the most represented, particularly in rural areas (35.05%) compared to urban & rural (30.84%) and urban (29.80%) datasets. The distribution of men and women is similar across all datasets, with women's proportion approximately 15% greater than that of men in all three datasets.

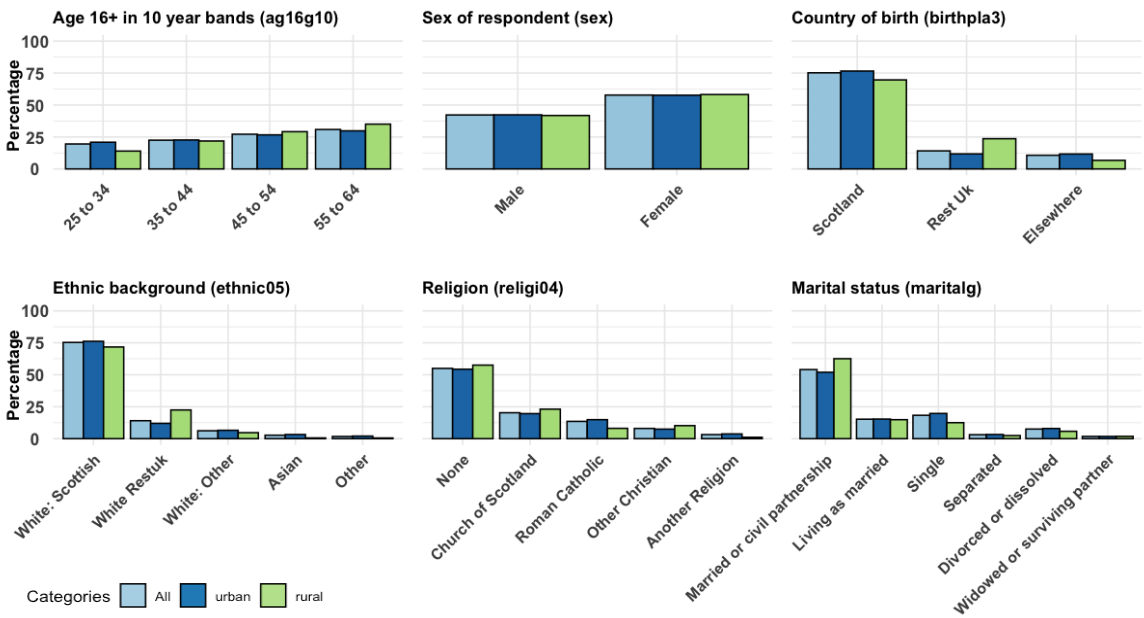


FIGURE 3.1 PROPORTION OF SOCIOECONOMIC VARIABLES ACROSS THE THREE DATASETS

Birthplace data indicated that around 70% of respondents were born in Scotland, making it the most predominant group across all datasets (Figure 3.1). However, the rural group had a lower percentage of individuals born in Scotland (69.64%) compared to the urban & rural (75.19%) and urban (76.56%) datasets. In contrast, rural areas had a higher percentage of individuals born in the rest of the UK (23.64%) compared to the urban & rural (14.12%) and urban (11.76%) datasets. Ethnic background data showed white Scottish individuals as predominant across all datasets, with the rural group showing a higher prevalence of white born in the rest of UK individuals (22.46%) compared to the urban & rural (14.06%) and urban (11.98%) datasets. The proportion of white individuals from outside the UK was similar in the urban & rural (6.19%) and urban (6.56%) datasets but about 2% lower in the rural dataset (4.69%). Asian and other ethnic groups had similar representations in the urban & rural and urban datasets (Asian: urban & rural = 2.73%, urban = 3.26%; Other: urban & rural = 1.74%, urban = 2.04%), but both groups had a very low presence in rural areas (Asian: 0.59%; Other: 0.54%).

Religion data indicated that over 50% of respondents did not profess religion, with rural areas showing the highest proportion (57.54%) (Figure 3.1). Among those who professed a religion, the Church of Scotland had a higher representation in rural areas (23.08%) compared to the urban & rural (20.30%) and urban (19.61%) datasets. In contrast, Roman Catholicism was more predominant in the urban & rural (13.53%) and urban (14.88%) datasets compared to rural areas (8.08%). Other religions had similar proportions in the urban & rural (3.19%) and urban (3.71%) datasets but a lower presence in rural areas (1.08%).

Regarding marital status, most respondents were married or in a civil partnership, with this status being about 10% higher in the rural group (62.57%) compared to the urban & rural (54.06%) and urban (51.95%) datasets (Figure 3.1). Singles were more common in the urban & rural (18.30%) and urban (19.73%) groups compared to

rural areas (12.52%). The proportion of widows was very low across all datasets (urban & rural = 1.78%, urban = 1.76%, rural = 1.85%).

Rural respondents reported higher life satisfaction (39.33%) compared to the urban & rural (33.72%) and urban (32.33%) datasets (Figure 3.2). The urban dataset had the highest percentage (36.61%) of respondents with the lowest life satisfaction (below mode). More than 70% of respondents across all datasets reported having 'very good' or 'good' general health, with rural respondents slightly higher at 77.96% compared to 74.46% urban & rural and 73.60% urban. This aligned with data on LTC, where the rural dataset reported fewer significant limitations (13.65%) compared to the urban & rural (16.22%) and urban (16.86%) datasets.

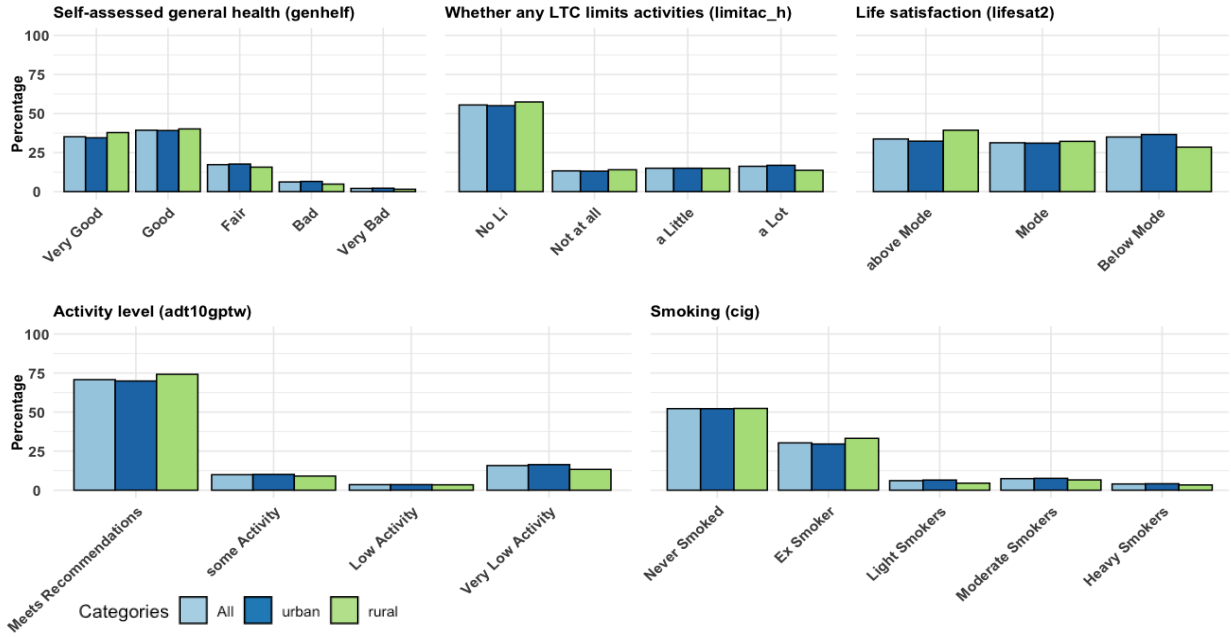


FIGURE 3.2 PROPORTION OF HEALTH-RELATED VARIABLES ACROSS THE THREE DATASETS

Most respondents in all datasets met recommended activity levels (adt10gptw) (Figure 3.2). Rural areas had a slightly lower percentage in the 'very low activity' category (rural = 13.31%, urban & rural = 15.77%, urban = 16.38%). Over 80% of respondents were never smokers or ex-smokers (never smoked: urban & rural = 52.22%, urban 52.19%, rural = 52.36%; ex-smoker: urban & rural = 30.30%, urban = 29.58%, rural = 33.20%). Heavy smokers comprised around 4% of the total (urban & rural = 3.98%, urban = 4.14%, rural = 3.36%).

A significant portion of the participants held a degree or higher qualification (Figure 3.3), with the urban & rural percentage at 43.21%, urban at 42.74% and rural at 45.08%. The majority fell into the II Managerial/Technical class category, which was more prevalent in the rural dataset (43.31%) compared to urban & rural (39.71%) and urban (38.82%). Employment status data indicated that around 75% of respondents were employed, with the highest proportion in the rural dataset (78.81%). The category 'ILO Unemployed', which referred to individuals who did not work during the reference week but were actively seeking and available for work according to the International Labour Organisation (ILO) definition (ILO, 1996), included less than 3% of cases across all datasets (urban & rural = 2.39%, urban = 2.55%, rural = 1.76%).

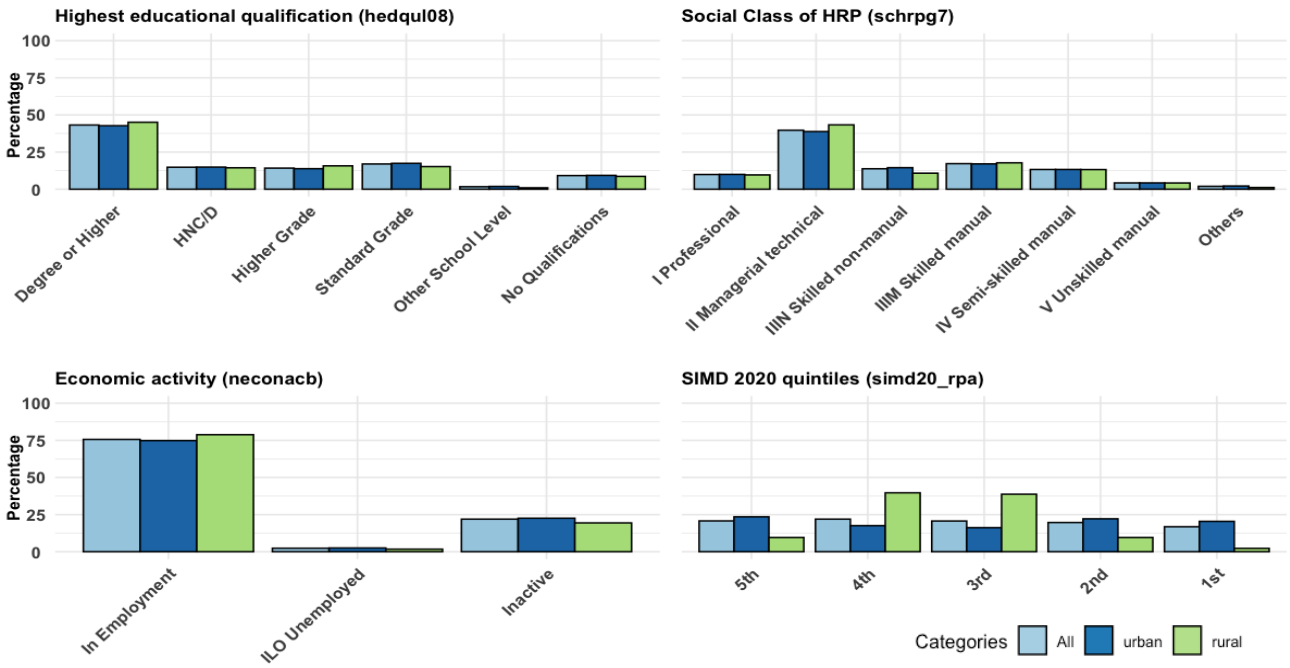


FIGURE 3.3 PROPORTION OF EDUCATIONAL AND SOCIOECONOMIC-RELATED VARIABLES ACROSS THE THREE DATASETS

The Scottish Index of Multiple Deprivation variable (simd20_rpa) (Figure 3.3), showed that the most deprived quintiles (1st and 2nd) and the least deprived quintile (5th) were more represented in the urban & rural (1st = 16.86%, 2nd = 19.70%, 5th = 20.78%) and urban (1st = 20.47%, 2nd = 22.19%, 5th = 23.54%) datasets compared to the rural dataset (1st = 2.30%, 2nd = 9.64%, 5th = 9.64%) (Figure 3.3). Conversely, the middle quintiles, third (3rd) and fourth (4th), were more represented in rural areas (3rd = 38.74%, 4th = 39.68%) compared to the urban & rural (3rd = 20.70%, 4th = 21.97%) and urban areas (3rd = 16.23%, 4th = 17.58%).

Concerning alcohol consumption (Figure 3.4), the percentage of individuals who exceeded the weekly limit for alcohol consumption was similar across the three groups: urban and rural combined, urban only, and rural only, at 24.02%, 24.02%, and 24.01%, respectively. The highest median alcohol units per week were in the rural dataset (4.67), slightly higher than the urban & rural (4.50) and urban (4.35) datasets. The interquartile range (IQR) for alcohol consumption was similar for the urban & rural (0.35- 13.57) and urban (0.33- 13.59) datasets but slightly narrower for the rural dataset (0.49- 13.50). The rural dataset also had the lowest maximum weekly consumption (302.50) compared to the urban & rural and urban datasets (412.50).

Wine was the most consumed alcoholic beverage weekly (Figure 3.4), with the median consumption in the rural dataset (0.69) almost double that of the urban & rural (0.35) and urban datasets (0.26). The urban dataset showed the narrowest range of wine consumption and the lowest maximum (126.00 vs. 297.00 for urban & rural, and rural). Spirits consumption was similar across all three groups, with a median of 0.23 units each; but the rural group had the lowest maximum consumption (154.00) compared to the urban & rural, and urban group (280.00). Regular beer is also similar across the three settings, with medians around 0.05. Strong beer, sherry, and alcopops consumption was very low across all three samples.

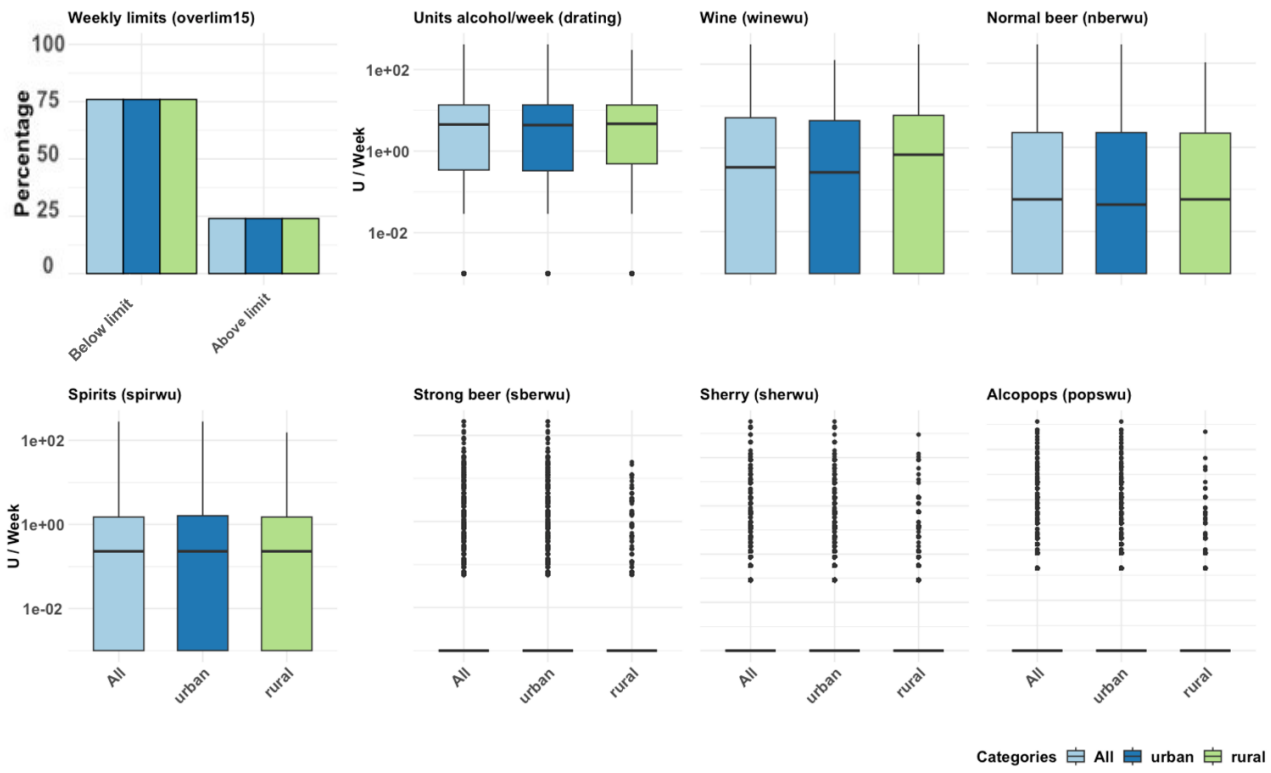


FIGURE 3.4 ALCOHOL CONSUMPTION ACROSS THE THREE DATASETS

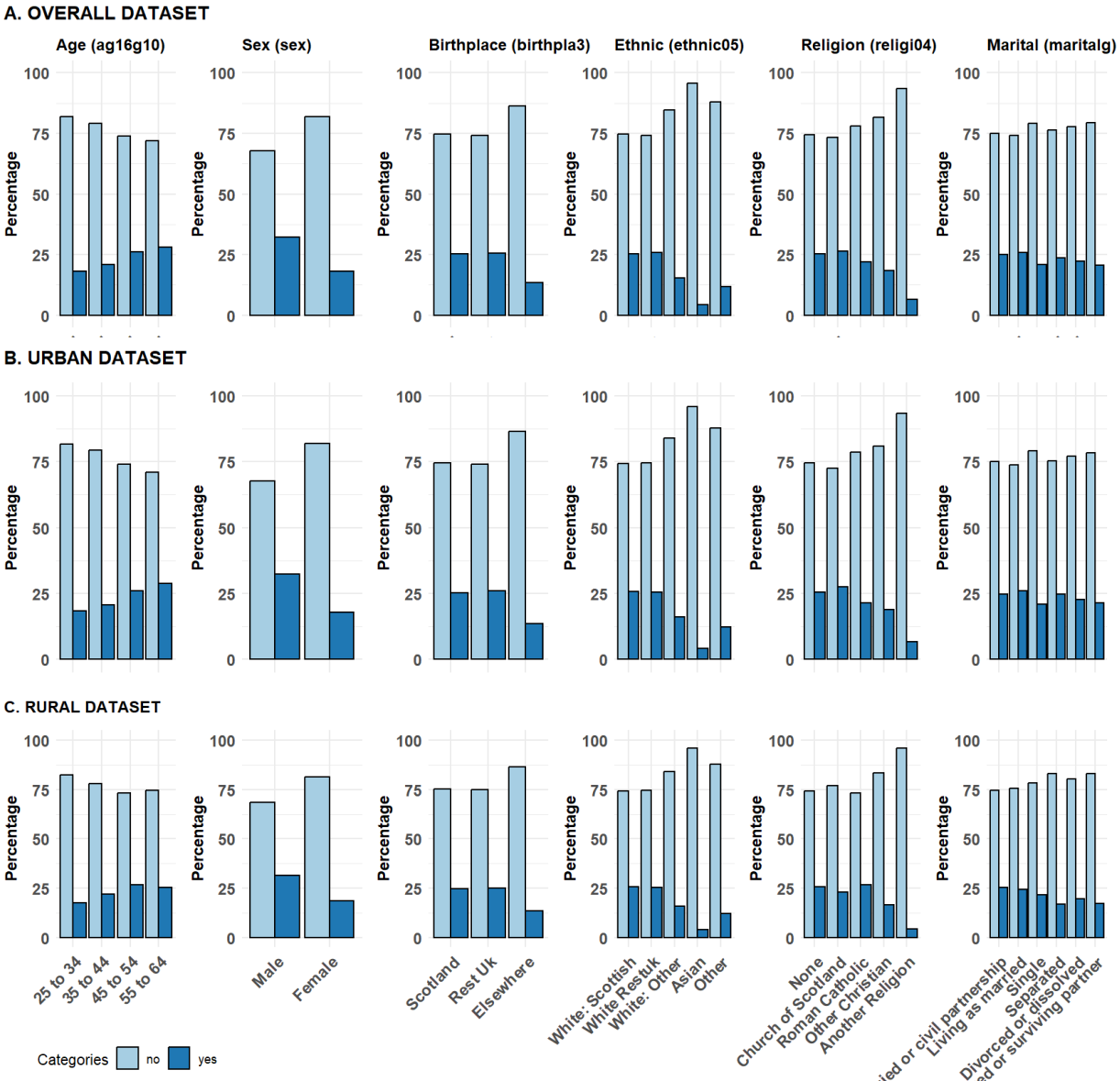
3.2 DISTRIBUTION OF THE OUTPUT VARIABLE IN THE URBAN & RURAL, URBAN AND RURAL DATASETS

The following plots illustrate the distribution of alcohol consumption below and above the 14-week recommended limit, represented by the output variable (overlim15), across the three datasets. A detailed breakdown of quantities and percentages is provided in Appendix Table I.2.

In all three datasets, most people across all age groups reported consuming alcohol below the weekly limit ('no') across the three datasets (Figure 3.5). However, the age group with more representation than those who reported consumption below the limits was the 25-34-year-old group (urban & rural = 81.81%, urban = 81.72% and rural = 82.35%). The group with more proportion in the above-the-limit category was 55-64 years old for the urban & rural (28.13%), and urban (28.90%) datasets and 45-54 (26.75%) for the rural dataset. The percentage of males consuming above the weekly limit was approximately 15% higher than that of females (males: urban & rural = 32.12%, urban = 32.30%, rural = 31.38%; females: urban & rural = 18.11%, urban = 17.95%, rural = 18.74%).

Both 'yes' and 'no' categories showed a similar pattern about birthplaces, ethnicities, and religion across the three datasets (Figure 3.5). About ten percent more of the participants born outside the UK reported consuming below the limit (urban & rural = 86.45%, urban = 86.44%, rural = 86.49%) compared with the two other groups (Scotland and Rest UK). Asians had the highest percentage in the below-limit category, with figures above 90% (urban & rural = 95.72%, urban = 95.88%, rural = 92.31%). Regarding the religions, the category 'another religion' had the lowest percentage in those cases classified above the limit (urban & rural = 6.48%, urban = 6.65%, rural = 4.17%).

With respect to marital status (Figure 3.5), the below-the-limit category was dominated by single and widowed individuals in the urban & rural (single = 79.01%, widowed = 79.40%), and urban (single = 79.13%, widowed = 78.48%) datasets and separated (83.02%) and widowed (82.93%) individuals in the rural dataset.On the other hand, living as married group was protagonist in the above-the-limit category in the urban & rural (25.75%) and urban (26.10%) datasets, and married or in civil partnerships group (25.29%) in rural dataset.



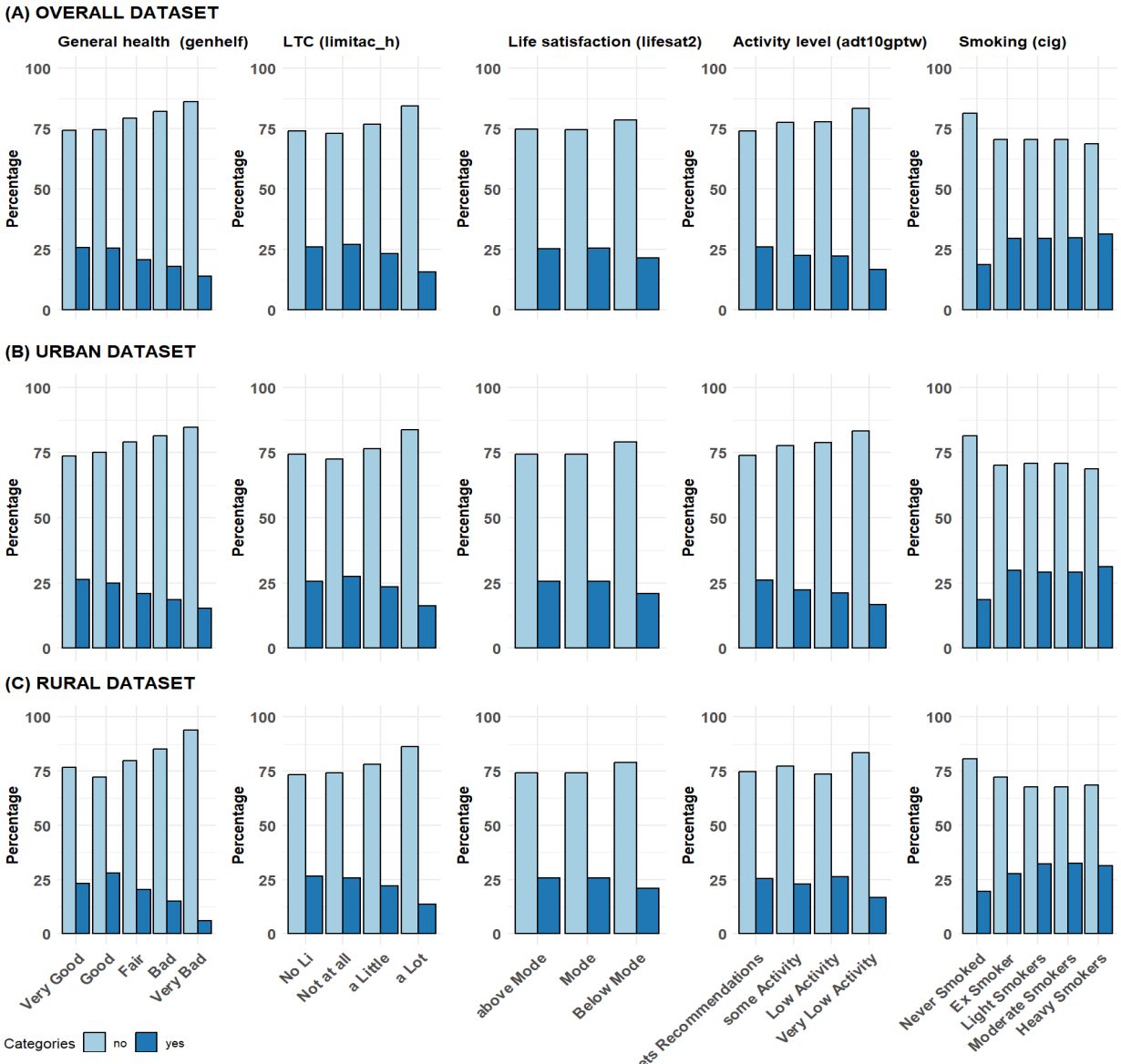
Categories: no = Below the weekly limit, yes = Above the

FIGURE 3.5 ALCOHOL CONSUMPTION RELATIVE TO WEEKLY LIMITS BY SOCIODEMOGRAPHIC VARIABLES ACROSS A. URBAN & RURAL DATASET (IN THIS PLOT AS OVERALL DATASET), B. URBAN DATASET AND C. RURAL DATASET

In relation to self-assessed general health (Figure 3.6), those who reported very bad health had a percentage higher than 80% in the category below the weekly limit, especially in the rural dataset (93.94%). The same trend was observed for those with an LTC that significantly limited their life, with percentages higher than 80% in the group below the limit across all three datasets (urban & rural = 84.26%, urban = 83.83%, rural = 86.38%). There were no big differences between the grade of life satisfaction and harmful alcohol consumption in the rural dataset (Figure 3.6), with values around 75% for the three categories of this variable (above mode, mode, and

below mode). However, in the urban & rural, and urban datasets, the group below mode (urban & rural = 78.49% and urban = 78.96%) showed numbers 4% higher than the rest of the groups for those who drank under the limit.

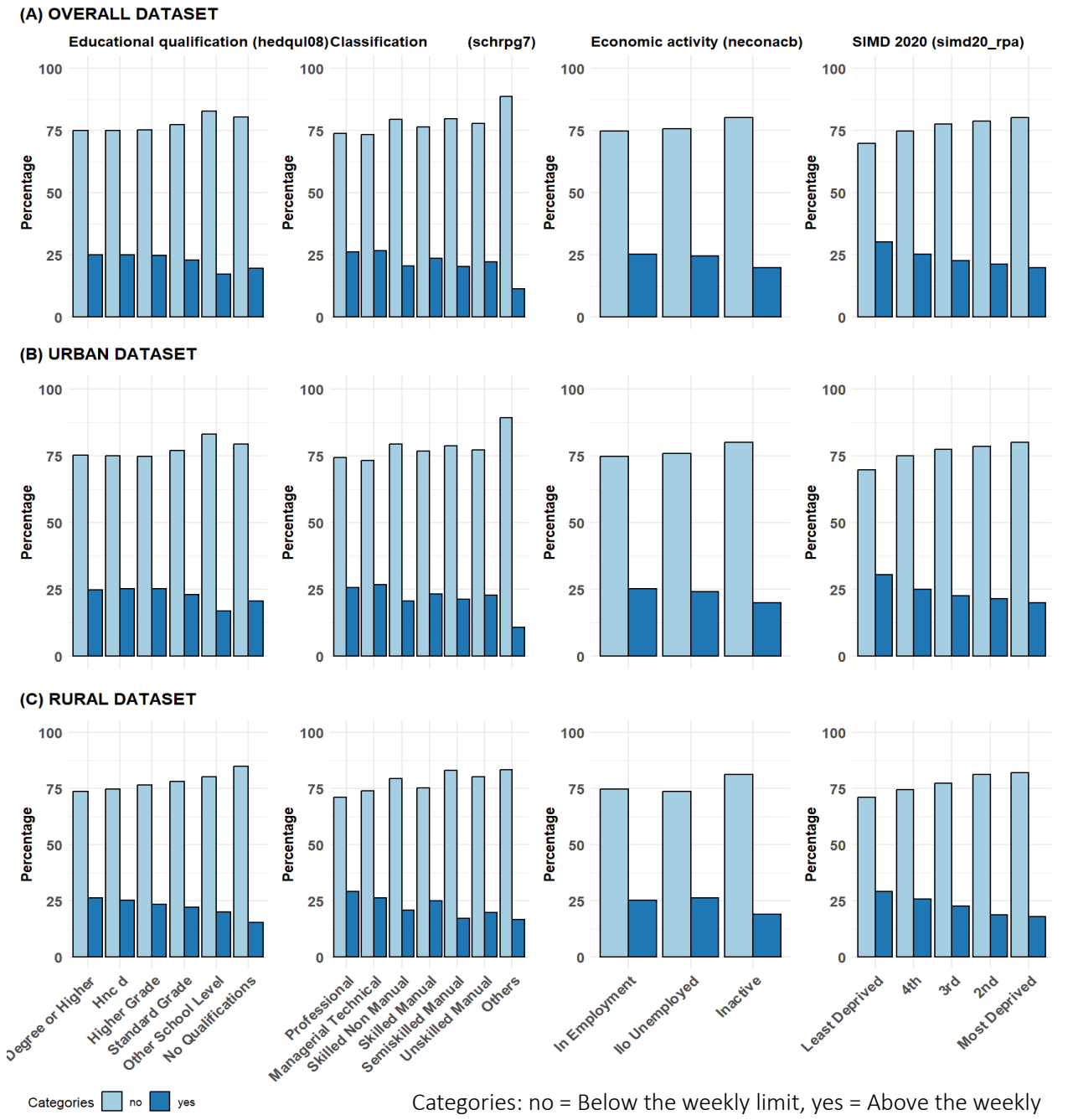
Individuals with very low activity levels also showed higher representation in the group below the limit (urban & rural = 83.29%, urban = 83.26%, rural = 83.39%). On the contrary, the group that met the recommendation had a higher percentage in the 'yes' category, mainly in the urban & rural (25.99%), and urban (26.15%) datasets. Regarding smoking habits, clear differences were seen between those who never smoked and the rest. More than 80% of the never-smoked group did not have harmful alcohol consumption (urban & rural = 81.23%, urban = 81.39%, rural = 80.59%). However, about one-third of heavy smokers exceeded the weekly alcohol consumption limit (urban & rural = 31.21%, urban = 31.15%, rural = 31.51%).



Categories: no = Below the weekly limit, yes = Above

FIGURE 3.6 ALCOHOL CONSUMPTION RELATIVE TO WEEKLY LIMITS BY HEALTH-RELATED VARIABLES ACROSS A. URBAN & RURAL DATASET (IN THIS PLOT AS OVERALL DATASET), B. URBAN DATASET AND C. RURAL DATASET

In three datasets, the groups with the higher educational levels (degree or higher, HNC/D, and higher grade) showed a higher percentage of respondents in the group above the limit (Figure 3.7), especially in the rural dataset (degree or higher = 26.31%, HNC/D = 25.32%, and higher grade = 23.41%). A similar pattern was observed with the social classification, where professionals (urban & rural = 26.25%, urban = 25.57%, rural = 29.05%) and managerial/technical (urban & rural = 26.62%, urban = 26.76%, rural = 26.14%) workers had the highest percentages in the group above the limit in all three datasets.

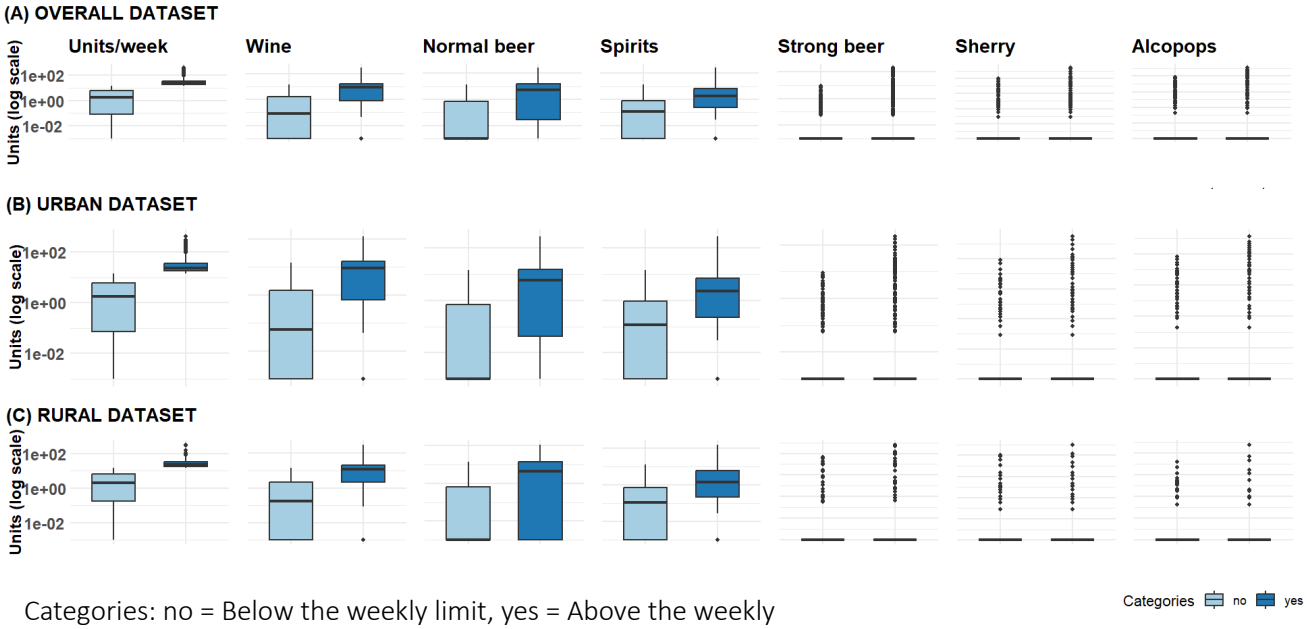


Categories: no = Below the weekly limit, yes = Above the weekly

FIGURE 3.7 ALCOHOL CONSUMPTION RELATIVE TO WEEKLY LIMITS BY EDUCATIONAL AND SOCIOECONOMIC-RELATED VARIABLES ACROSS A. URBAN & RURAL DATASET (IN THIS PLOT AS OVERALL DATASET), B. URBAN DATASET AND C. RURAL DATASET

The variable related to employment (Economic activity) showed a similar distribution across the three datasets (Figure 3.7), with those employed (urban & rural, urban and rural = 25.24%) or looking for a job (ILO Unemployed) (urban & rural = 24.43%, urban = 24.11%, rural = 26.32%) having more representation in the group above the limit. Regarding deprivation (SIMD), those least deprived also had the highest percentages in the harmful alcohol consumption group across all three datasets (urban & rural = 30.23%, urban = 30.35%, rural = 29.11%), exceeding by around 10% those in the most deprived segment (urban & rural = 19.80%, urban = 19.85%, and rural = 18.00%).

As expected, the total alcohol units consumed weekly (Figure 3.8) was higher in the above-limit group [Median (IQR) in units: urban & rural = 23.07(17.91-34.10), urban = 23.43(17.85-34.50) and rural = 22.40 (18.00-33.08)]. The beverage consumed in the highest units in the group above the limits was wine [Median (IQR) in units: urban & rural = 9.00 (0.84-15.75), urban = 9.00(0.69-15.75), rural = 12.00 (2.25-20.81)], followed by regular beer [Median (IQR) in units: urban & rural = 5.08 (0.03-15.23), urban = 6.00(0.04-15.23), rural = 4.35 (0.00-14.00)] and spirits [Median (IQR) in units: urban & rural = 1.88 (0.23-7.00), urban = 2.25(0.23-7.00), rural = 1.50 (0.23-6.00)], in all three datasets (Figure 3.8). Wine was also the most consumed beverage in the group below the limit, but the second was spirits, and regular beer followed it. The consumption of other beverages was very low in both groups, below and above the limit.



Categories: no = Below the weekly limit, yes = Above the weekly

FIGURE 3.8 ALCOHOL CONSUMPTION BY BEVERAGE TYPE ACROSS A. URBAN & RURAL DATASET (IN THIS PLOT AS OVERALL DATASET), B. URBAN DATASET AND C. RURAL DATASET

Overall, the group with weekly alcohol consumption below the limits, had a protagonist presence, and the patterns were very similar in the three datasets. The higher incidence of auto-reported above-limit alcohol consumption was among older age groups, males, individuals with a very good or good general health condition, without LTC, with life satisfaction above or equal to the mode, smokers, and individuals with higher educational levels, employed and belonging to the least deprived areas. The beverage most consumed for them

was wine. This suggested that certain socio-demographic factors and lifestyle choices were strongly associated with higher alcohol consumption.

3.3 RESULTING DATASETS AFTER DATA SPLITTING

After splitting the datasets (see Section 2.8), each subset resulted in an approximate 70:30 ratio of training to test cases across the urban & rural, urban-only, and rural-only datasets. The urban & rural training and test datasets had the highest total case count, while the rural training and test subsets had the fewest cases overall. The dimensions of each dataset after splitting are detailed in Table 3.1.

TABLE 3.1 DIMENSIONS OF TRAINING AND TEST SUBSETS

Dataset		Before split	After split[a]			
			Train subset		Test subset	
Urban & Rural	dw	11,185	dw.train	7,831 (70.01%)	dw.test	3,354 (29.99%)
Urban	dwu	8,965	dwu.train	6,277 (70.02%)	dwu.test	2,688 (29.98%)
Rural	dwr	2,220	dwr.train	1,555 (70.06%)	dwr.test	665(29.94%)

[a] The percentages are calculated for the dimensions of the datasets before the split.

3.4 MISSING DATA ANALYSIS AND RESULTING DATASETS AFTER DELETION

There were 588 missing values and 10,778 (96.36%) complete cases in the urban & rural dataset (dw). Five of the 18 variables in the initial dataset did not have missing data: urban-rural classification (urbrur_all), Survey Year (syear), SIMD quintiles (SIMD20_RPa), Sex (sex) and Age (ag16g10) (Figure 3.9). The remaining variables' missing data percentage was at most 2%.

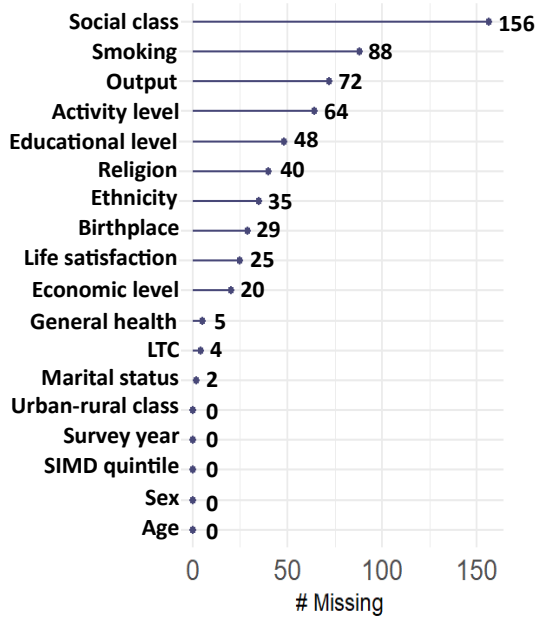


FIGURE 3.9 MISSING DATA PER VARIABLE

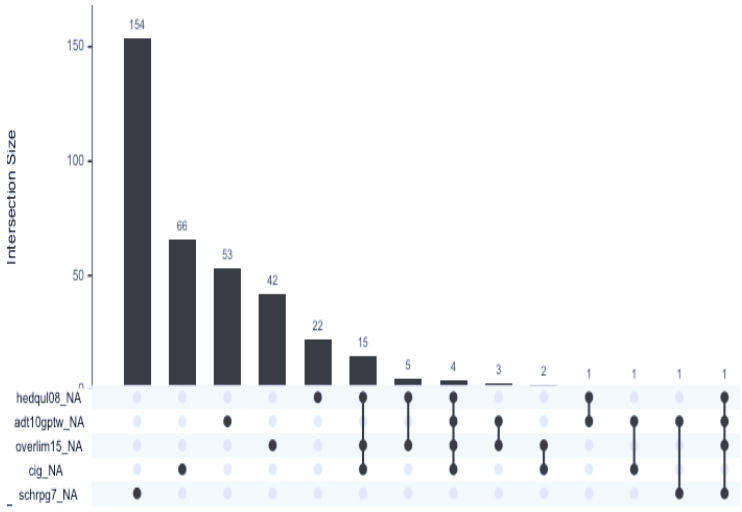


FIGURE 3.10 INTERSECTION OF MISSING DATA

The variable with the most missing data was Socio-economic Classification (schpg7), with 156 missing values (1.39%). Of these, 154 cases had missing data only in schpg7, and two cases had missing data in schpg7 as well as in Physical Activity level (adt10gptw) and another in educational levels (hedqul08) (Figure 3.10). The next variable with significant missing data was Smoking (cig), with 88 missing cases. Among these, 66 cases had

missing data only in cig, and 15 cases had missing data in cig as well as in overlim15, and hedqul08. Four cases had missing data in four variables: in cig, overlim15, adt10gptw, and hedqul08. However, most missing data cases involved only one variable. The missing data in the variables religion (religi04), ethnicity (ethnic05), birthplace (birthpla3), life satisfaction (lifesat2), economic activity (neconacb), general health (genhelf), LTC (limitac_h), and marital status (maritalg) did not have interactions with other variables; the missing data was specific to these variables.

After listwise deletion (see Section 2.9), cases were removed from each dataset, with deletions remaining under 4% across all training and test datasets. Specifically, the urban & rural training dataset (dw.train) lost 283 cases (3.61%), the urban training dataset (dwu.train) lost 241 cases (3.84%), and the rural training dataset (dwr.train) lost 57 cases (3.67%). In the test datasets, 124 cases (3.70%) were removed from the urban & rural test subset (dw.test), 95 cases (3.53%) from the urban test subset (dwu.test), and 14 cases (2.10%) from the rural test subset (dwr.test). The urban training subset had the most deleted cases (3.84%), while the rural test subset, the smallest dataset, had the lowest (2.10%). These results were summarised in Table 3.2.

TABLE 3.2 DIMENSIONS OF TRAINING AND TEST SUBSETS BEFORE AND AFTER LISTWISE DELETION

Dataset	Before deletion		After deletion[a]	
	Train subset	Test subset	Train subset	Test subset
Urban & rural (dw)	7,831	3,354	7,548 (96.39%)	3,230 (96.30%)
Urban (dwu)	6,277	2,688	6,036 (96.16%)	2,593 (96.47%)
Rural (dwr)	1,555	665	1,498 (96.33%)	651 (97.90%)

[a] The percentages are calculated concerning the dimensions of the subsets before deletion.

3.5 RESULTING DATASETS AFTER OVERSAMPLING

After applying oversampling (see 2.10.1) to the urban & rural (dw.train), urban (dwu.train), and rural (dwr.train) training datasets with different ratios (30%, 40%, and 50%), nine new datasets were generated: three from each of the original training datasets (dw.train, dwu.train and dwr.train). Table 3.3 summarises the resulting datasets.

TABLE 3.3 DATASETS OBTAINED BY OVERSAMPLING

Original dataset	Datasets created by oversampling			
	Name	Output Variable: overlim15[a]		Total number of cases
		Majority class (No)	Minority class (Yes)	
dw.train	up30_dw	5,741 (70%)	2,461 (30%)	8,202
	up40_dw	5,741 (60%)	3,828 (40%)	9,569
	up50_dw	5,741 (50%)	5,741 (50%)	11,482
dwu.train	up30_dwu	4,587 (70%)	1,966 (30%)	6,553
	up40_dwu	4,587 (60%)	3,059 (40%)	7,646
	up50_dwu	4,587 (50%)	4,587 (50%)	9,174
dwr.train	up30_dwr	1,138 (70%)	488 (30%)	1,626
	up40_dwr	1,138 (60%)	759 (40%)	1,897
	up50_dwr	1,138 (50%)	1,138 (50%)	2,276

[a] The percentages are calculated for the total number of cases.

The final sizes of these datasets varied depending on the oversampling ratio applied, with the 50:50 ratio, which achieves an equal balance between majority and minority classes, producing the largest datasets. The maximum sizes reached were 11,482 cases for dw.train, 9,174 cases for dwu.train, and 2,276 cases for dwr.train.

3.6 RESULTING DATASETS AFTER SMOTE

The application of SMOTE to the urban & rural (dw.train), urban (dwu.train), and rural (dwr.train) datasets resulted in nine modified datasets for each, generating a total of 27 datasets (Table 3.4). These datasets exhibited varying class distributions depending on the resampling ratios and the nearest-neighbour parameter (k). Specifically, three values of k (3, 5, and 10) and three resampling ratios (0.5, 0.75, and 1.0) were used, which determined the final dataset sizes and class balance (see Section 2.10.2).

TABLE 3.4 DATASETS OBTAINED BY SMOTE

Original				Datasets created by oversampling		
Dataset	Name	k	or	Output Variable: overlim15		Total cases
				Majority class (No) [a]	Minority class (Yes) [a]	
dw.train	smk3r50_dw	3	0.5	5,741 (66.67%)	2,870 (33.33%)	8,611
	smk5r50_dw	5				
	smk10r50_dw	10				
	smk3r75_dw	3	0.75	5,741 (57.15%)	4,305 (42.85%)	10,046
	smk5r75_dw	5				
	smk10r75_dw	10				
	smk3r100_dw	3	1	5,741 (50%)	5,741 (50%)	11,482
	smk5r100_dw	5				
	smk10r100_dw	10				
dwu.train	smk3r50_dwu	3	0.5	4,587 (66.67%)	2,293 (33.33%)	6,880
	smk5r50_dwu	5				
	smk10r50_dwu	10				
	smk3r75_dwu	3	0.75	4,587 (75.15%)	3,440 (42.85%)	8,027
	smk5r75_dwu	5				
	smk10r75_dwu	10				
	smk3r100_dwu	3	1	4,587 (50%)	4,587 (50%)	9,174
	smk5r100_dwu	5				
	smk10r100_dwu	10				
dwr.train	smk3r50_dwr	3	0.5	1,138 (66.67%)	569 (33.33%)	1,707
	smk5r50_dwr	5				
	smk10r50_dwr	10				
	smk3r75_dwr	3	0.75	1,138 (57.16%)	853 (42.84%)	1,991
	smk5r75_dwr	5				
	smk10r75_dwr	10				
	smk3r100_dwr	3	1	1,138 (50%)	1,138 (50%)	2,276
	smk5r100_dwr	5				
	smk10r100_dwr	10				

[a] The percentages are calculated for the total number of cases.

The largest and most balanced datasets were achieved using a 1:1 resampling ratio, resulting in dataset sizes of up to 11,482 cases for dw.train, 9,174 cases for dwu.train, and 2,276 cases for dwr.train. In contrast, smaller datasets with a dominant majority class were obtained using the lowest resampling ratio (0.5). The choice of k also influenced the distribution of synthetic samples, with higher values of k leading to a more spread-out distribution of the minority class.

3.7 CHI-SQUARED

The residuals plot in Figure 3.11 showed the standardized residuals, which represented the differences between the observed and expected frequencies under the null hypothesis. The plot revealed that the 'Yes' category from overlim15 (alcohol consumption above the 14-unit weekly limit) and the 'rural' category from urbrur_all (living in a rural area) had the largest positive residuals (0.1). This suggests that there was a slight overrepresentation of rural individuals consuming alcohol above the 14-unit limit. However, the residuals for other categories were small, indicating that the observed and expected frequencies aligned closely for those groups.

In Figure 3.12, the contribution of each cell to the chi-squared statistic is shown. Although the contribution of the 'rural' category from urbrur_all and the 'Yes' category from overlim15 was higher than others, contributing 97.57 to the total chi-squared value, the overall chi-squared statistic (0.011) was very low, and the p-value (0.915) exceeded the threshold of 0.05, indicating that the result was not statistically significant. Therefore, despite the observed deviations, there was no strong evidence to suggest an association between living in an urban or rural area and consuming alcohol above or below the 14-unit weekly limit.

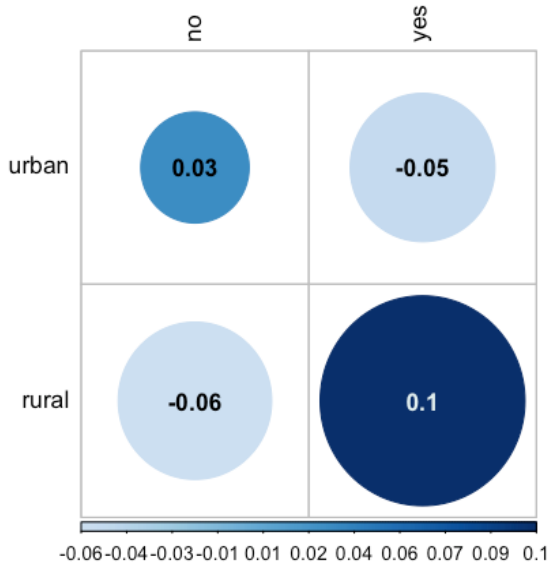


FIGURE 3.11 STANDARDISED RESIDUALS PLOT

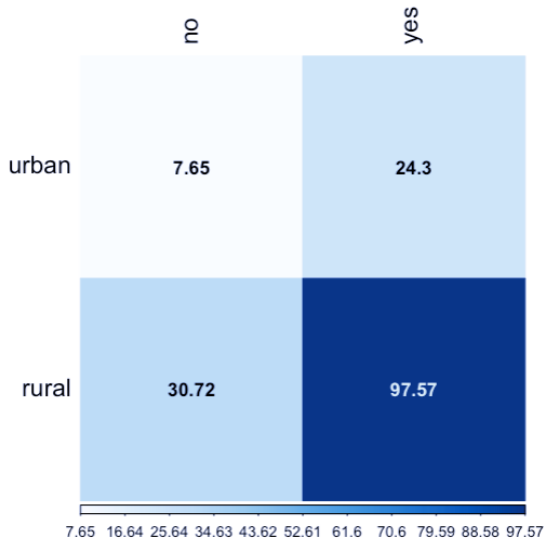


FIGURE 3.12 CHI-SQUARED CONTRIBUTION PLOT

3.8 URBAN & RURAL DATASETS

The original urban & rural dataset (dw) contained 11,185 cases and 18 variables, including 17 predictors and the outcome variable, overlim15. After the data split (Section 2.8), 7,831 cases were assigned to the urban & rural training subset (dw.train) and 3,354 to the test subset (dw.test) (Table 3.1). Following this, listwise deletion was applied to handle missing data (Section 2.9), resulting in 7,548 cases in dw.train and 3,230 cases in dw.test (Table 3.2).

To address the class imbalance in the training urban & rural subset (Section 2.10), oversampling (Section 2.10.1) and SMOTE (Section 2.10.2) were applied using different parameters. This process produced 12 distinct urban & rural training datasets (three from the oversampling and nine from SMOTE) with varying dimensions (Tables 3.3 and 3.4). The smallest dataset, with 8,202 cases, was obtained when oversampling was applied with a 70:30 proportion (70% for the majority class and 30% for the minority class). In contrast, the largest dataset, with 11,482 cases, was generated using SMOTE with or = 1 and k = 10. Each dataset included a mix of cases from both urban and rural contexts.

The following sections present the results of the logistic regression (LR) modelling with stepwise selection, applied to all 13 urban & rural training datasets (the original dataset, three datasets generated through oversampling, and nine generated via SMOTE). Additionally, the evaluation and validation of the selected model are presented. This analysis aimed to enhance predictive performance and identify the key predictors of alcohol consumption exceeding the recommended weekly limit of 14 units, focusing on the overall population without distinguishing between urban and rural residency.

3.8.1 LR IN URBAN & RURAL DATASETS

After applying stepwise selection to all 13 urban & rural training datasets, the model with the fewest variables was the one obtained from the original urban & rural training dataset (dw.train), containing 11 predictors (Table 3.5). In contrast, the model that included the most variables, a total of 16, was obtained using the smk5r100_dw dataset, generated with SMOTE, where k = 5 and or = 1. The remaining models derived from the datasets created with SMOTE included the next highest number of variables, with 15 variables in each. All models generated from the urban & rural training datasets created through oversampling included 13 variables after stepwise selection.

TABLE 3.5 URBAN & RURAL DATASETS: STEPWISE SELECTION RESULTS

Variables	Original	Oversampling			SMOTE								
					k = 3			k = 5			k = 10		
		up30	up40	up50	or=0.5	or=0.75	or=1	or=0.5	or=0.75	or=1	or=0.5	or=0.75	or=1
1. Survey year (syear)					✓	✓	✓	✓	✓	✓	✓	✓	✓
2. Age (ag16g10)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3. Sex (sex)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4. Birthplace (birthpla3)										✓			
5. Ethnicity (ethnic05)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
6. Religion (religi04)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
7. Marital status (maritalg)		✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
8. General health (genhelf)		✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
9. LTC (limitac_h)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
10. Life-satisfaction (lifesat2)													
11. Activity level (adt10gptw)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
12. Smoking (cig)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
13. Educational level (hedqul08)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
14. Social class (schrpg7)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
15. Economic activity (neconacb)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
16. SIMD quintile (simd20_rpa)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
17. Urban-rural class (urbrur_all)			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

The variable related to self-reported life satisfaction (lifesat2) was absent from all models selected through stepwise selection across the urban & rural datasets. In contrast, the following variables were consistently included in all selected models:

- Age (ag16g10)
- Sex (sex)
- Ethnicity (ethnic05)
- Religion (religi04)
- LTC (limitac_h)
- Activity level (adt10gptw)
- Smoking (cig)
- Educational level (hedqul08)
- Social class (schrpg7)
- Economy activity (neconacb)
- SIMD quintile (simd20_rpa)

The variable Year (syear), which was not selected in either the original dataset or the oversampled datasets, appeared in all models generated from the SMOTE-created datasets. Similarly, the variable urban-rural classification (urbrur_all), which was not included in the model from the original training dataset (dw.train), appeared in all models generated from the SMOTE-created datasets and in those from the oversampled training datasets, except for the up30_dw dataset (ratio 70:30).

3.8.2 CROSS-VALIDATION IN URBAN & RURAL MODELS

Table 3.6 summarises the cross-validation results, showcasing the performance metrics (AUC ROC, sensitivity, specificity, and accuracy) for all the urban & rural models assessed in the training datasets. All the estimates, statistics and p values, can be reviewed in the Appendix (Table I.3 to I.15).

TABLE 3.6 URBAN & RURAL MODELS: CROSSVALIDATION RESULTS

Datasets				Metrics (%)			
				ROC	Sensitivity	Specificity	Accuracy
Original urban & rural training dataset			dw.train	68.63%	97.57%	7.72%	76.06%
Oversampling urban & rural training datasets		70:30	up30_dw	68.99%	92.86%	19.02%	70.70%
		60:40	up40_dw	69.01%	79.86%	42.11%	64.76%
		50:50	up50_dw	69.35%	62.45%	66.90%	64.67%
SMOTE urban & rural training datasets	k = 3	or = 0.5	smk3r50_dw	71.24%	87.95%	32.20%	69.37%
		or = 0.75	smk3r75_dw	72.85%	75.46%	54.50%	66.48%
		or = 1	smk3r100_dw	73.74%	64.94%	69.31%	67.13%
	k = 5	or = 0.5	smk5r50_dw	71.22%	88.27%	32.79%	69.78%
		or = 0.75	smk5r75_dw	72.71%	75.39%	53.65%	66.07%
		or = 1	smk5r100_dw	74.46%	65.19%	70.36%	67.78%
	k = 10	or = 0.5	smk10r50_dw	71.46%	88.20%	32.50%	69.63%
		or = 0.75	smk10r75_dw	73.24%	76.10%	55.72%	67.37%
		or = 1	smk10r100_dw	74.83%	66.11%	70.61%	68.36%

The worst performance was observed with the original urban & rural training dataset (dw.train), which had the lowest AUC ROC (68.63%) and specificity (7.72%), despite its high sensitivity (97.57%) and moderate accuracy (76.06%). This highlights the challenge of class imbalance affecting specificity.

With the oversampled urban & rural datasets, overall metrics improved, particularly sensitivity. Among these, the model using a 50:50 oversampling ratio (up50_dw) achieved the highest AUC ROC (69.35%) and demonstrated a better balance across metrics, with superior specificity (66.90%) and accuracy (64.67%) compared to the 70:30 (up30_dw) and 60:40 (up40_dw) ratios, despite lower sensitivity (62.45%).

The SMOTE datasets further improved performance. The model smk10r100_dw (k = 10, or = 1), achieved the highest AUC ROC (74.83%) and balanced metrics, including sensitivity (66.11%), specificity (70.61%), and accuracy (68.36%). Similarly, the smk5r100_dw model (k = 5, or = 1) performed well, with an AUC ROC of 74.46%, sensitivity of 65.19%, specificity of 70.36%, and accuracy of 67.78%. Among models with k = 3, smk3r100_dw (or = 1) had the best performance, achieving an AUC ROC of 73.74%, sensitivity of 64.94%, specificity of 69.31%, and accuracy of 67.13%.

Overall, the smk10r100_dw model (SMOTE k = 10 or = 1) stands out with its high AUC ROC and balanced sensitivity and specificity, highlighting SMOTE with a higher k and optimal oversampling ratio as the best approach for addressing class imbalance.

3.8.3 FITTING AND INTERPRETING THE URBAN & RURAL MODEL

The urban & rural LR model included 15 variables, detailed in Table 3.7. The table outlines all the categories for each variable.

TABLE 3.7 URBAN & RURAL MODEL: PREDICTORS AND THEIR CATEGORIES

Variable	Categories
1. Survey year (syear)	▪ *2017* ▪ 2018 ▪ 2019 ▪ 2021
2. Age (ag16g10)	▪ 25-34 ▪ 35-44 ▪ 45-54 ▪ 55-64
3. Sex (sex)	▪ *Male* ▪ Female
4. Ethnicity (ethnic05)	▪ *White: Scottish* ▪ White: rest of the UK ▪ White: Other ▪ Other (other minority ethnics including Asians)
5. Religion (religi04)	▪ *None* ▪ Church of Scotland ▪ Roman Catholic ▪ Other (other Christian or another religion)
6. Marital status (maritalg)	▪ *Married/civil partnership* ▪ Living as married ▪ Single ▪ Separated or Widowed (or divorced or dissolved)
7. General health (genhelf)	▪ *Very good* ▪ Good ▪ Fair ▪ Bad-Very bad
8. LTC (limitac_h)	▪ *No limitations* ▪ Not at all ▪ A little ▪ A lot
9. Activity level (adt10gptw)	▪ *Meets recommendations* ▪ Some activity ▪ Low activity ▪ Very low activity
10. Smoking (cig)	▪ *Never smoked* ▪ Ex-smoker ▪ Light ▪ Moderate ▪ Heavy

11. Educational level (hedqul08)	▪ *Degree or higher* ▪ HNC/D ▪ Higher (higher grade or equivalent) ▪ School grade (standard grade or equivalent) ▪ No qualifications
12. Social class (schrpg7)	▪ *Professional* ▪ Managerial technical ▪ Skilled non-manual ▪ Skilled manual ▪ Semi-skilled manual ▪ Unskilled manual-Others
13. Economic activity (neconacb)	▪ *In employment* ▪ Unemployed-Inactive
14. SIMD quintile (simd20_rp)	▪ *Least deprived* ▪ 4th ▪ 3rd ▪ 2nd ▪ Most deprived
15. Urban-rural class (urbrur_all)	▪ *Urban* ▪ Rural

The ORs and 95% CIs for each category within the variables are displayed in Figure 3.13. The reference category for each variable, indicated in italics in Table 3.7, was used for comparison.

There is a significant increase in the OR for each subsequent year compared to 2017, with 2021 exhibiting the highest OR (2.142, CI = 1.887- 2.433). Sex is a critical predictor, with women demonstrating significantly lower odds (OR = 0.456, CI = 0.420- 0.496) of exceeding the weekly alcohol consumption limit compared to men. The likelihood of exceeding this limit also increases with age, with the 55-64 age group having the highest OR (2.225, CI = 1.933- 2.561). Ethnicity and religion are also influential factors. Individuals identified as 'other' ethnicity (OR = 0.376, CI = 0.315- 0.449) and those affiliated with 'other religions' (OR = 0.399, CI = 0.335- 0.474) have significantly lower odds compared to the White: Scotland and None religion groups, respectively. Additionally, individuals identifying as Roman Catholic have significantly lower odds (OR = 0.775, CI = 0.678- 0.885) than those with no religious affiliation. Marital status is another significant determinant. Single individuals (OR = 0.837, CI = 0.738- 0.950) and those who are separated, divorced, or widowed (OR = 0.724, CI = 0.627- 0.836) show lower odds than those who are married or in a marriage-like relationship.

Long-term conditions (LTC) and physical activity levels also influence the outcome, with individuals engaging in very low physical activity (category: 'Very low activity', OR = 0.582, CI = 0.503- 0.673) and those severely affected by an LTC (category: 'A lot', OR = 0.540, CI = 0.448- 0.651) having lower odds compared to their respective reference groups (no limitation and meeting physical activity recommendations, respectively). Smoking status is a notable predictor, with smokers having significantly higher odds of exceeding the weekly alcohol consumption limit compared to those who have never smoked. Heavy smokers have the highest odds (OR = 2.558, CI = 2.035- 3.215).

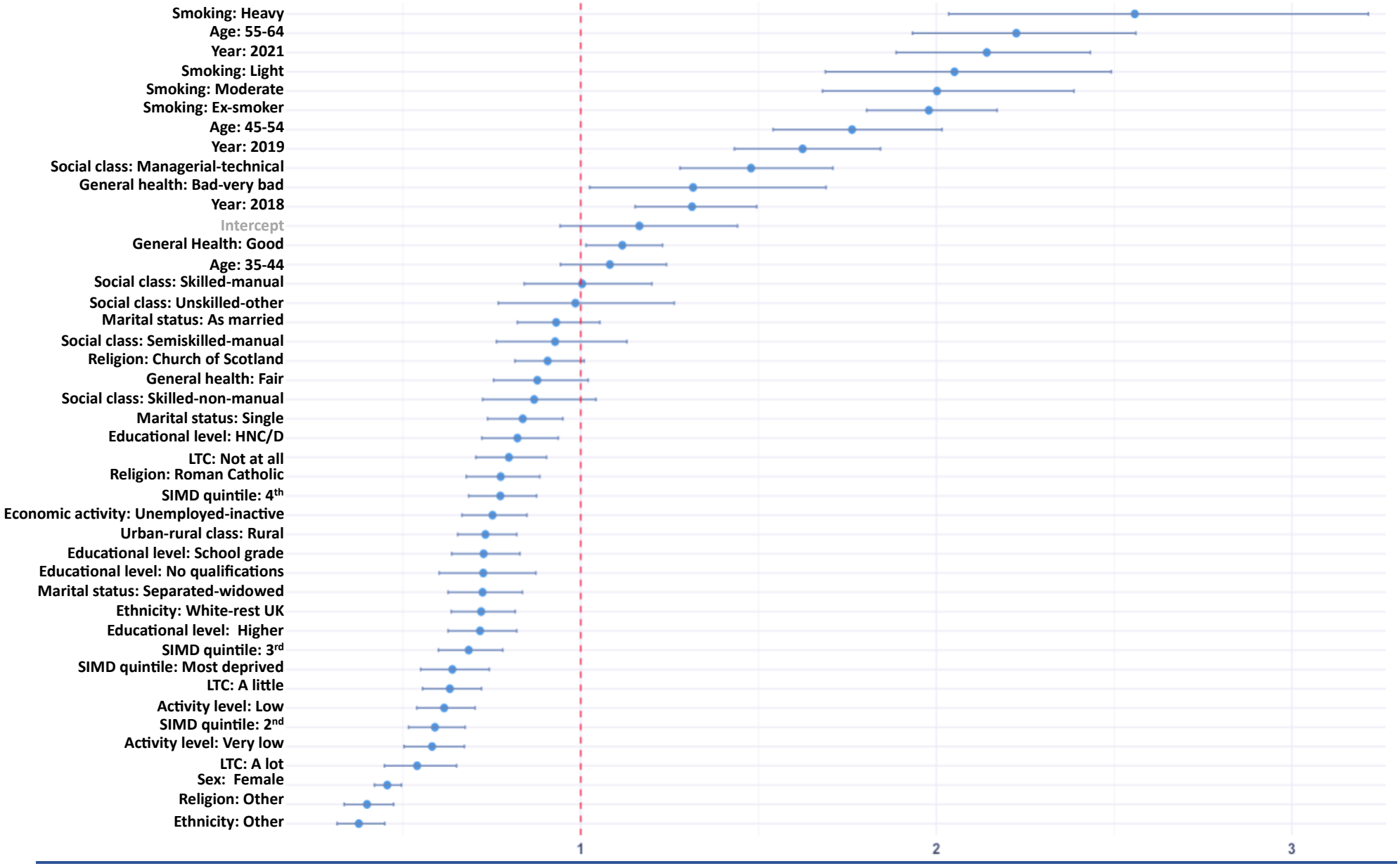


FIGURE 3.13 URBAN & RURAL MODEL: ODDS RATIOS AND 95% CONFIDENCE INTERVALS

Educational attainment is strongly associated with the odds of exceeding the alcohol consumption limit. Lower educational qualifications are associated with significantly lower odds compared to having a degree, with individuals holding a 'Higher Grade' qualification showing the lowest odds ratio (OR = 0.717, CI = 0.627–0.820), though this is very close to the other categories. Social class also affects the outcome, with individuals in the 'II Managerial technical' class showing higher odds (OR = 1.479, CI = 1.279- 1.709) compared to those in the 'I Professional' class. Increased deprivation is correlated with lower odds of exceeding the alcohol consumption limit, with individuals in the '2nd quintile' having the lowest OR (0.590, CI = 0.516- 0.675). Furthermore, individuals residing in rural areas have lower odds (OR = 0.732, CI = 0.654- 0.820) than urban residents. Employment status is also a significant factor, with 'ILO unemployed & Inactive' individuals having lower odds (OR = 0.752, CI = 0.666- 0.849) than those in employment.

The category 35–44 in Age is a non-significant factor compared with the reference group 25–34 (p = 0.261). Likewise, in Religion (religi04), the category Church of Scotland is non-significant compared with None (p = 0.076). In Marital Status (maritalg), the category As married shows no significant difference compared with Married (p = 0.258). Similarly, for Self-assessed General Health (genhelf2), the category Fair is non-significant compared with Very good (p = 0.090). Finally, in the Social Class variable (schrpg7), the categories Skilled non-manual (p = 0.131), Skilled manual (p = 0.962), Semi-skilled manual (p = 0.458), and Unskilled manual-other (p = 0.903) are all non-significant compared with the Professional category.

In summary, the analysis reveals that sex, age, ethnicity, religion, marital status, health (long-term conditions), educational level, employment status, deprivation level, urban-rural residency, smoking status, and physical activity level, are all significant predictors of exceeding the weekly alcohol consumption limit.

3.8.4 EVALUATING THE URBAN & RURAL MODEL IN THE TEST SUBSET

The performance metrics in Table 3.8 demonstrate the evaluation of the LR model selected and applied to the urban & rural test datasets.

TABLE 3.8 URBAN & RURAL MODEL: METRICS IN THE TEST SUBSET

AUC ROC	Sensitivity	Specificity	Accuracy	PPV	F1-score
65.80%	66.93%	56.88%	59.29%	32.85%	44.07%

The model can moderately discriminate between positive and negative cases (AUC = 65.80%) but does not reach 70% to be considered good performance (Figure 3.14). The sensitivity (recall) is 66.93%. However, the specificity is very low: the model correctly identifies only 56.88% of the true negative cases. The accuracy (59.29%) indicates that just over half of the total predictions (both positive and negative) made by the model are correct. Therefore, the model's overall performance is only slightly better than random guessing (50%). The PPV (precision) is particularly low; when the model predicts a positive case, it is correct only 32.85% of the time. The low specificity and PPV suggest a high rate of false positives, which can reduce the model's reliability in predicting positive cases. The F1-score (44.07%), influenced by the low PPV value, is also low, indicating that the model is ineffective in balancing the trade-off between capturing positive cases and minimising false positives.

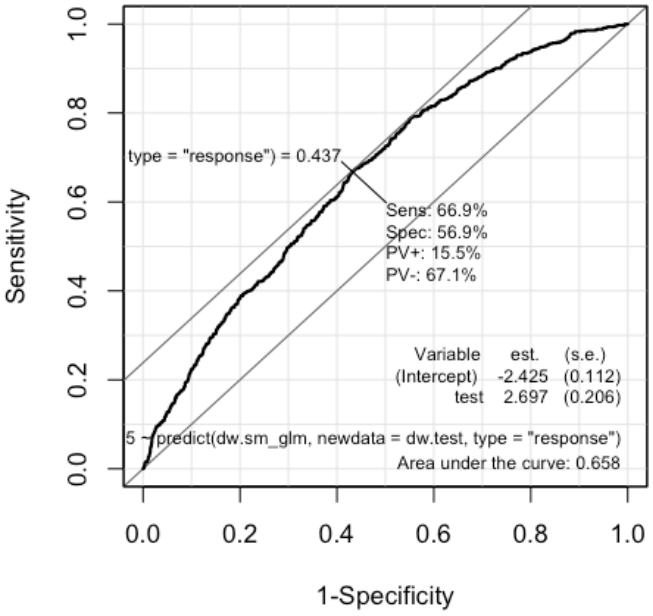


FIGURE 3.14 URBAN & RURAL TEST SUBSET: MODEL'S EVALUATION RESULTS

3.9 URBAN DATASETS

The original urban dataset (dwu) comprises 17 columns (16 predictors and the outcome variable: overlim15) and 8,965 cases, all classified as urban according to the Scottish Government Urban Rural Classification (Geographic Information Science & Analysis Team, 2014, 2018, 2022). After splitting the data (Section 2.8) into training and test subsets and applying oversampling (Section 2.10.1) and SMOTE (Section 2.10.2) to the training subset, using various parameters to address class imbalance (Section 2.10), 13 urban training datasets were generated, each with different dimensions (Tables 3.3 and 3.4). These datasets contain cases only from urban contexts.

This section details the outcomes of applying logistic regression (LR) with stepwise selection to the urban training datasets: the original dataset, three created through oversampling, and nine generated using SMOTE. The goal of the analysis was to enhance predictive accuracy and pinpoint the main factors contributing to alcohol consumption exceeding the recommended weekly limit of 14 units for the urban group. The section also covers the evaluation and validation of the selected model in the test subset.

3.9.1 LR IN URBAN DATASETS

Most models obtained by stepwise selection from the 13 urban training datasets incorporate 12 variables (Table 3.9). Three models derived from the data generated by SMOTE include 14 variables [smk3r75_dwu (k = 3, or = 0.75), smk3r100_dwu (k = 3, or = 1), smk10r100_dwu (k = 10, or = 1)], the maximum number of variables

among the models. The model derived from the original training data (dwu.train) (without oversampling or SMOTE) includes the fewest variables, with a minimum of 10.

TABLE 3.9 URBAN DATASETS: STEPWISE SELECTION RESULTS

Variables	Original	Oversampling			SMOTE								
					k = 3			k = 5			k = 10		
		up30	up40	up50	or=0.5	or=0.75	or=1	or=0.5	or=0.75	or=1	or=0.5	or=0.75	or=1
1. Survey year (syear)					✓	✓	✓	✓	✓	✓	✓	✓	✓
2. Age (ag16g10)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3. Sex (sex)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4. Birthplace (birthpla3)			✓	✓	✓	✓	✓			✓			✓
5. Ethnicity (ethnic05)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
6. Religion (religi04)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
7. Marital status (maritalg)		✓		✓							✓		✓
8. General health (genhelf)			✓			✓	✓		✓	✓		✓	✓
9. LTC (limitac_h)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
10. Life-satisfaction (lifesat2)	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	
11. Activity level (adt10gptw)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
12. Smoking (cig)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
13. Educational level (hedqul08)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
14. Social class (schrpg7)		✓				✓	✓	✓	✓	✓	✓	✓	✓
15. Economic activity (neconacb)													
16. SIMD quintile (simd20_rpa)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

The variables listed below are included in all the models:

- Age (ag16g10)
- Sex (sex)
- Ethnicity (ethnic05)
- Religion (religi04)
- LTC (limitac_h)
- Activity level (adt10gptw)
- Smoking (cig)
- Educational level (hedqul08)
- SIMD quintile (simd20_rpa)

Similar to the training datasets corresponding to cases not divided between urban and rural areas (dw), the variable Survey year (syear) is included in all models obtained through stepwise selection from the training datasets created by SMOTE. However, it is not included in the models derived from the original training data or the training data created by oversampling.

3.9.2 CROSS-VALIDATION IN URBAN MODELS

Table 3.10 provides a summary of the cross-validation results, highlighting the performance metrics (ROC, sensitivity, specificity, and accuracy) for all the urban models evaluated in the training datasets. All estimates, statistics, and p-values can be found in the Appendix (Tables I.16 to I.28).

TABLE 3.10 URBAN MODELS: CROSSVALIDATION RESULTS

Datasets				Metrics (%)			
				ROC	Sensitivity	Specificity	Accuracy
Original urban training dataset			dwu.train	69.32%	97.37%	9.07%	76.17%
Oversampling urban training datasets		70:30	up30_dwu	69.52%	92.48%	21.05%	71.05%
		60:40	up40_dwu	70.23%	80.24%	45.40%	66.30%
		50:50	up50_dwu	70.41%	63.10%	66.34%	64.72%
SMOTE urban training datasets	k = 3	or = 0.5	smk3r50_dwu	71.11%	88.44%	32.12%	69.67%
		or = 0.75	smk3r75_dwu	73.07%	75.66%	55.27%	66.92%
		or = 1	smk3r100_dwu	73.57%	65.40%	69.24%	67.32%
	k = 5	or = 0.5	smk5r50_dwu	71.09%	88.24%	31.56%	69.35%
		or = 0.75	smk5r75_dwu	73.32%	75.70%	55.33%	66.97%
		or = 1	smk5r100_dwu	73.89%	65.44%	69.53%	67.49%
	k = 10	or = 0.5	smk10r50_dwu	71.74%	87.95%	34.27%	70.06%
		or = 0.75	smk10r75_dwu	73.23%	76.01%	55.88%	67.38%
		or = 1	smk10r100_dwu	74.30%	65.76%	69.54%	67.65%

The original training dataset showed the worst performance among the urban training datasets (Table 3.10). However, models generated using oversampling methods showed improvements. The best model from the oversampled datasets is the one with a 50:50 oversampling ratio (50% majority class and 50% minority class) (up50_dwu). This model achieved the highest AUC ROC value (70.41%) among the three oversampled urban training datasets. While its sensitivity is lower than the other models (up50_dwu = 63.10% vs up30_dwu = 92.48% and up40_dwu = 80.24%), it is more balanced in terms of specificity (66.34%) and accuracy (64.72%).

For the SMOTE-based models, the model derived from the smk10r100_dwu dataset (k = 10 or = 1) exhibited the highest AUC ROC (74.30%) and specificity (69.54%). The highest sensitivity (88.44%) was achieved by the smk3r50_dwu model (k = 3 or = 0.5), while the highest accuracy (70.06%) was seen with the smk10r50_dwu model (k = 10 or = 0.5). However, both of these models displayed very low specificity (smk3r50_dwu = 32.12% and smk10r50_dwu = 34.27%).

It can be concluded that the most balanced model is the one obtained from smk10r100_dwu (k = 10, or = 1), which, as mentioned, had the highest AUC ROC (74.30%) and specificity (69.54%) values while also showing more balanced sensitivity (65.76%) and accuracy (67.65%) in relation to the other metrics.

3.9.3 FITTING AND INTERPRETING THE URBAN MODEL

The model chosen for the urban population group contains the 14 variables detailed in Table 3.11. The reference category for each variable is indicated in italics.

TABLE 3.11 URBAN MODEL: PREDICTORS AND THEIR CATEGORIES

Variable	Categories
1. Survey year (syear)	▪ *2017* ▪ 2018 ▪ 2019 ▪ 2021
2. Age (ag16g10)	▪ 25-34 ▪ 35-44 ▪ 45-54 ▪ 55-64
3. Sex (sex)	▪ *Male* ▪ Female
4. Birthplace (birthpla3)	▪ *Scotland* ▪ Rest of the UK ▪ Elsewhere
5. Ethnicity (ethnic05)	▪ *White: Scottish* ▪ White: rest of the UK ▪ White: Other ▪ Other (other minority ethnics including Asians)
6. Religion (religi04)	▪ *None* ▪ Church of Scotland ▪ Roman Catholic ▪ Other (other Christian or another religion)
7. Marital status (maritalg)	▪ *Married/civil partnership* ▪ Living as married ▪ Single ▪ Separated or Widowed (or divorced or dissolved)
8. General health (genhelf)	▪ *Very good* ▪ Good ▪ Fair ▪ Bad-Very bad
9. LTC (limitac_h)	▪ *No limitations* ▪ Not at all ▪ A little ▪ A lot
10. Activity level (adt10gptw)	▪ *Meets recommendations* ▪ Some activity ▪ Low activity ▪ Very low activity
11. Smoking (cig)	▪ *Never smoked* ▪ Ex-smoker ▪ Light ▪ Moderate ▪ Heavy
12. Educational level (hedqul08)	▪ *Degree or higher* ▪ HNC/D ▪ Higher (higher grade or equivalent)

	▪ School grade (standard grade or equivalent) ▪ No qualifications
13. Social class (schrpg7)	▪ *Professional* ▪ Managerial technical ▪ Skilled non-manual ▪ Skilled manual ▪ Semi-skilled manual ▪ Unskilled manual-Others
14. SIMD quintile (simd20_rp)	▪ *Least deprived* ▪ 4th ▪ 3rd ▪ 2nd ▪ Most deprived

The Figure 3.15 summarises the estimates, coefficients, OR, and 95% CI from a LR analysis of the urban dataset. The interpretation of the model obtained from the urban cases is very similar to that of the urban & rural data. Year 2021 is associated with increased odds of the outcome compared to 2017 (the reference year), with an odds ratio (OR) of 1.875 (CI: 1.630–2.156). Age groups 45–54 and 55–64 show higher odds of the outcome, with ORs of 2.034 (CI: 1.750–2.365) and 2.490 (CI: 2.125–2.916), respectively. Females have significantly lower odds (OR = 0.424, CI: 0.386–0.465) than males, suggesting a gender disparity already observed in previous population groups concerning harmful alcohol consumption.

Individuals of other ethnicities, not from the UK, have significantly lower odds of consuming above the limit (OR = 0.365, CI: 0.263–0.506). Regarding religion, those in the 'Other religion' category have the lowest OR (0.477, CI: 0.393–0.578) compared with the rest of the categories in this variable.

Marital status showed close ORs. Considering 'Married' as the reference category, the ORs of the rest of the categories range from a minimum of 0.841 (CI: 0.719–0.984) for 'Separated or Widowed' to a maximum of 1.081 (CI: 0.939–1.246) for the 'As Married' category.

Individuals reporting 'Fair' health have the lowest odds for the general health variable (genhelf2) with an OR of 0.735 (CI: 0.623–0.866). However, those reporting 'Good' health (OR = 0.909, CI: 0.815–1.014) or 'Bad/Very Bad' health (OR = 1.089, CI: 0.840–1.413) showed slight differences between them, as well as in comparison with those who considered their health to be 'Very Good,' which was the reference category.

Additionally, individuals with a significant degree of limitation (LTC variable) have lower odds of the outcome than the rest (OR = 0.589, CI: 0.483–0.719). People with low (OR = 0.638, CI: 0.552–0.738) and very low activity levels (OR = 0.576, CI: 0.490–0.676) have significantly lower odds than those meeting the physical activity requirements.

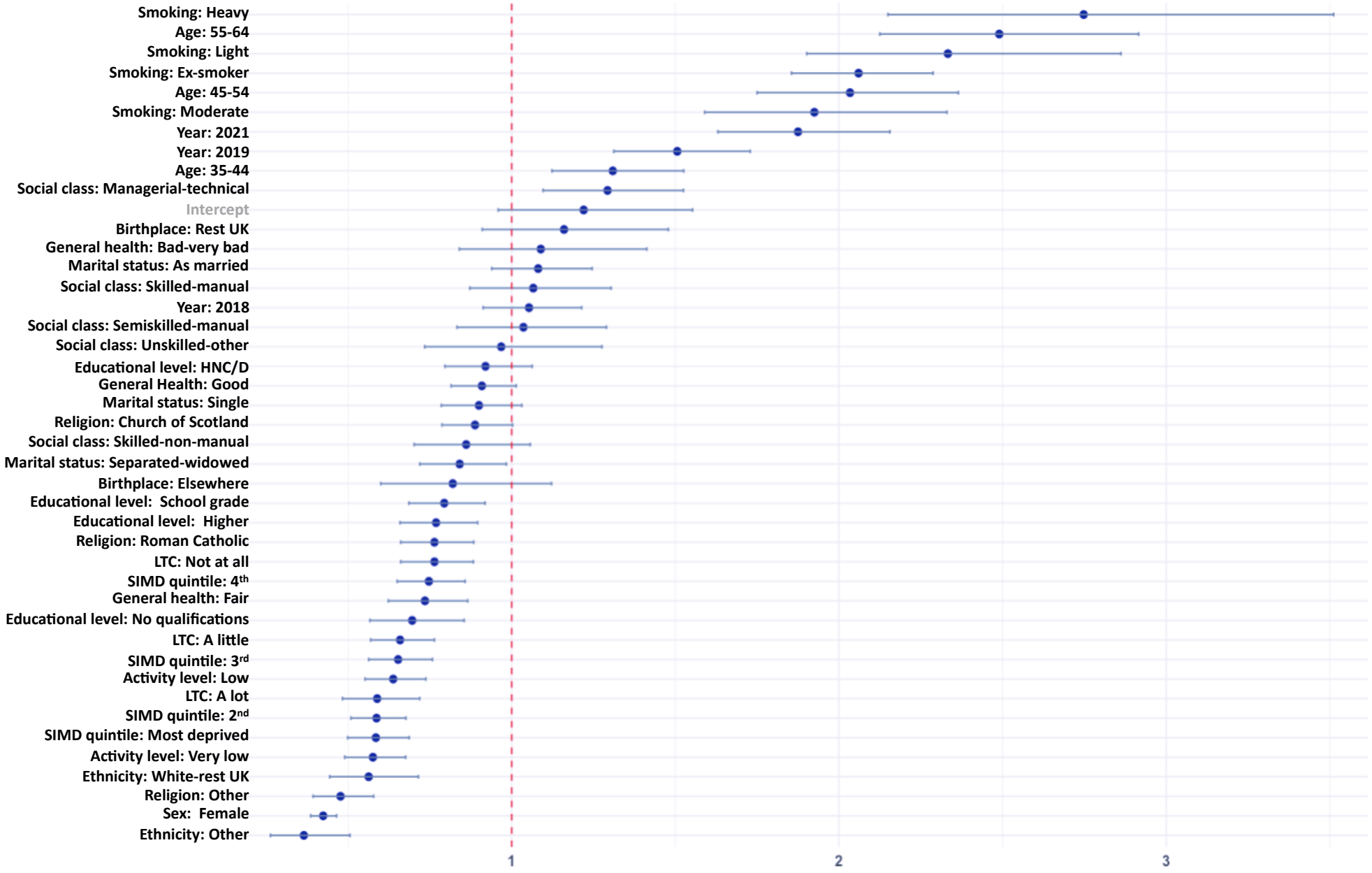


FIGURE 3.15 URBAN MODEL: ODDS RATIOS AND 95% CONFIDENCE INTERVALS

In relation to education level (hedqul08), individuals with no qualifications have significantly lower odds (OR = 0.696, CI: 0.567–0.855) compared to those with a degree or HNC/D. The OR close to 1 indicates that there is little difference between those with an HND/D and a degree in relation to the probability of consuming alcohol above the 14-weekly limit recommended.

From a socio-economic perspective, managerial and technical class individuals (schrpg7) have higher odds (OR = 1.293, CI: 1.096–1.525) than those in the professional class. However, skilled manual (OR = 1.066, CI: 0.872–1.304) and semi-skilled manual (OR = 1.036, CI: 0.833–1.290) groups have a similar probability to the professional class of exceeding the weekly alcohol limit consumption.

Increasing levels of deprivation (simd20_rpa) are associated with progressively lower odds of the outcome, with the most deprived quintile showing an OR of 0.585 (CI: 0.499–0.687) and the second most deprived an OR of 0.587 (CI: 0.509–0.677) compared to the least deprived quintile.

The year 2018 (p = 0.477) is a non-significant factor compared with 2017, the reference group. Similarly, in the birthplace variable (birthpla3), the categories 'Rest of UK' (p = 0.230) and 'Elsewhere' (p = 0.215) are non-significant compared with 'Scotland' and in religion variable the category 'Church of Scotland' in comparison with 'none' religion. In Marital Status (maritalg), the category 'As Married' and 'Single' showed no significant difference compared with 'Married' (p = 0.278 and p = 0.128, respectively). Similarly, for General Health (genhelf2), the category 'Good' and 'Bad-Very bad' are non-significant compared with 'Very Good' (p = 0.088 and p = 0.520, respectively). In 'Educational level' the category HNC/D (p = 0.259) with 'Professional'. Finally, in the Social Class variable (schrpg7), the categories 'Skilled non-manual' (p = 0.153), 'Skilled manual' (p = 0.531), 'Semi-skilled manual' (p = 0.749), and 'Unskilled manual-other' (p = 0.818) are all non-significant compared with the 'Professional' category.

Overall, these findings highlight the significant impact of sociodemographic characteristics (age, sex, birthplace, ethnicity, religion and marital status), health status (general health and LTC), lifestyle behaviours (smoking and activity level), and socio-economic (education level and social class) factors on the observed outcome in the urban group.

3.9.4 EVALUATING THE URBAN MODEL IN THE TEST SUBSET

The performance metrics for the LR model applied to the urban cases are shown in Table 3.12.

TABLE 3.12 URBAN MODEL: METRICS IN THE TEST SUBSET

ROC	Sensitivity	Specificity	Accuracy	PPV	F1-score
66.90%	73.17%	52.43%	57.35%	32.35%	44.87%

The model's performance is not adequate. The ROC is below 70% (66.90%), which indicates that it does not have a good discrimination capacity between cases above and below the weekly alcohol consumption limit (Figure 3.16). The sensitivity (73.17%), as in the general dataset, is high but is not balanced with the specificity (52.43%), which is very low; therefore, the model does not correctly identify the negative cases, leading to a Type I error. The total correct predictions (Accuracy = 57.35%) are only slightly above 50%. This shows a performance improvement of only around 7% compared to if the predictions were made at random. The PPV, as in the case of the dataset without the urban-rural classification distinction, is extremely low, with only 32.35% of the positive predictions made by the model being correct. This has consequences on the F1-score (44.87%), with a low value as well, suggesting that although the model is quite good at identifying positive cases, its accuracy is compromised by the high rate of false positives.

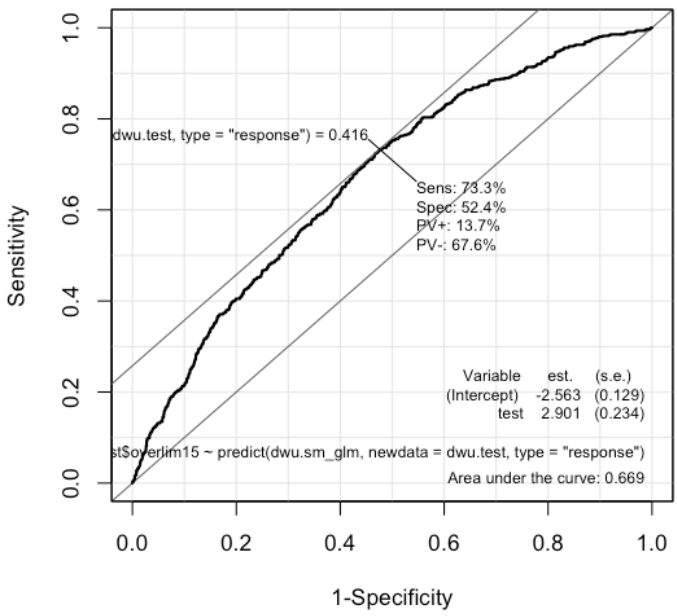


FIGURE 3.16 URBAN TEST SUBSET: MODEL'S EVALUATION RESULTS

3.10 RURAL DATASETS

The rural dataset (dwr) was initially created by filtering the rural cases from the combined urban & rural dataset based on the Scottish Government Urban Rural Classification (Geographic Information Science & Analysis Team, 2014, 2018, 2022). It included 2,220 observations and 17 variables: 16 predictors and the outcome variable (overlim15). After partitioning the data into training and test subsets (Section 2.8), oversampling (Section 2.10.1) and SMOTE (Section 2.10.2) were applied to the training subset to address class imbalance (Section 2.10), generating 13 different rural training datasets with varying dimensions (Tables 3.3 and 3.4).

This section details the outcomes of applying LR with stepwise selection to these rural training datasets. The datasets included the original, three oversampled versions, and nine generated using SMOTE. The goal was to improve model accuracy and identify the primary factors contributing to alcohol consumption exceeding the recommended 14-unit weekly limit for the rural population. Additionally, the section discusses how the selected model was evaluated and its performance assessed using the test subset.

3.10.1 LR IN RURAL DATASETS

Table 3.13 presents the variables included in each of the 13 LR models, for the rural cases, depending on the datasets selected for the LR stepwise selection.

TABLE 3.13 RURAL DATASETS: STEPWISE SELECTION RESULTS

Variables	Original	Oversampling			SMOTE								
					k = 3			k = 5			k = 10		
		up30	up40	up50	or=0.5	or=0.75	or=1	or=0.5	or=0.75	or=1	or=0.5	or=0.75	or=1
1. Survey's year (syear)						✓	✓	✓	✓	✓		✓	✓
2. Age (ag16g10)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3. Sex (sex)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4. Birthplace (birthpla3)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
5. Ethnicity (ethnic05)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
6. Religion (religi04)	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
7. Marital Status (maritalg)			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
8. Life satisfaction (lifesat2)													
9. General Health (genhelf)			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
10. LTC (limitac_h)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
11. Activity level (adt10gptw)							✓			✓			
12. Smoking (cig)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
13. Educational level (hedqul08)	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
14. Social class (schrpg7)		✓						✓	✓	✓		✓	✓
15. Economic activity (neconacb)		✓				✓	✓		✓			✓	✓
16. SIMD quintile (simd20_rpa)		✓		✓			✓		✓	✓			✓

The model selected from the original data (dwr) has the least number of variables among all datasets: eight variables. The models from the other datasets have between nine and 14 variables. The maximum of 14 variables is seen in the datasets generated by SMOTE: smk3r100_dwr (k = 3, or = 1), smk5r75_dwr (k = 5, or = 0.75), smk5r100_dwr (k = 5, or = 1), and smk10r100_dwr (k = 10, or = 1).

In all models generated through stepwise selection, the following variables were invariably included:

- Age (age16g10)
- Sex (sex)
- Birthplace (birthpla3)
- Ethnicity (ethnic05)
- LTC (limitac_h)
- Smoking (cig)

The variable Survey year (syear) is included in most of the models obtained through stepwise selection from the datasets generated by SMOTE, except in smk3r50_dwr (k = 3, or = 0.5) and smk10r50_dwr (k = 10, or = 0.5).

3.10.2 CROSS-VALIDATION IN RURAL MODELS

Table 3.14 summarises the performance metrics for all 13 models generated through stepwise selection, highlighting the impact of the different rural datasets on the model outcomes. Detailed estimates, statistics, and p-values are available in the Appendix (Tables I.29 to I.41).

TABLE 3.14 RURAL MODELS: CROSSVALIDATION RESULTS

Datasets				Metrics (%)			
				ROC	Sensitivity	Specificity	Accuracy
Original rural training dataset			dwr.train	66.27%	97.42%	8.44%	76.04%
Oversampling rural training datasets		70:30	up30_dwr	69.42%	90.90%	22.70%	70.43%
		60:40	up40_dwr	68.29%	77.29%	42.17%	63.24%
		50:50	up50_dwr	69.64%	62.57%	65.90%	64.23%
SMOTE rural training datasets	k = 3	or = 0.5	smk3r50_dwr	70.37%	85.94%	32.62%	68.17%
		or = 0.75	smk3r75_dwr	71.64%	74.45%	57.66%	67.25%
		or = 1	smk3r100_dwr	73.37%	63.97%	71.51%	67.74%
	k = 5	or = 0.5	smk5r50_dwr	69.35%	85.59%	31.46%	67.55%
		or = 0.75	smk5r75_dwr	73.08%	74.45%	57.23%	67.07%
		or = 1	smk5r100_dwr	74.81%	66.26%	72.14%	69.20%
	k = 10	or = 0.5	smk10r50_dwr	70.24%	85.38%	29.81%	66.86%
		or = 0.75	smk10r75_dwr	73.55%	74.08%	59.04%	67.63%
		or = 1	smk10r100_dwr	74.28%	65.50%	71.07%	68.29%

As in the datasets already analysed, in the rural datasets, the original dataset had the worst performance: with a ROC below 70% (66.27) and a very unbalanced sensitivity-specificity (97.42%- 8.44% respectively) and accuracy of 76.44%. In the oversampled dataset, the completely balanced one (up50_dwr) had the best performance because it had the highest values of AUC ROC (69.64%) among the three oversampled datasets, with a balance between sensitivity (62.57%), specificity (65.90%), and accuracy (64.23%). The models generated from the SMOTE datasets generally showed better performance, and one of them had the best performance of all models tested: the smk5r100_dwr, with an AUC ROC of 74.81%, sensitivity of 66.26%, specificity of 72.14%, and accuracy of 69.20%.

3.10.3 FITTING AND INTERPRETING THE RURAL MODEL

The model chosen for the rural dataset was that obtained from the SMOTE dataset smk5r100_dwu (k = 5, or = 1). Table 3.15 details the variables contained in the model. The first category, which is also italicised, serves as the reference category for the calculation of the ORs.

TABLE 3.15 RURAL MODEL: PREDITORS AND THEIR CATEGORIES

Variable	Categories
1. Survey year (syear)	▪ *2017* ▪ 2018 ▪ 2019 ▪ 2021

2. Age (ag16g10)	▪ 25-34 ▪ 35-44 ▪ 45-54 ▪ 55-64
3. Sex (sex)	▪ *Male* ▪ Female
4. Birthplace (birthpla3)	▪ *Scotland* ▪ Rest of the UK ▪ Elsewhere
5. Ethnicity (ethnic05)	▪ *White: Scottish* ▪ White: rest of the UK ▪ White: Other ▪ Other (other minority ethnics including Asians)
6. Religion (religi04)	▪ *None* ▪ Church of Scotland ▪ Roman Catholic ▪ Other (other Christian or another religion)
7. Marital status (maritalg)	▪ *Married/civil partnership* ▪ Living as married ▪ Single ▪ Separated or Widowed (or divorced or dissolved)
8. General health (genhelf)	▪ *Very good* ▪ Good ▪ Fair ▪ Bad-Very bad
9. LTC (limitac_h)	▪ *No limitations* ▪ Not at all ▪ A little ▪ A lot
10. Activity level (adt10gptw)	▪ *Meets recommendations* ▪ Some activity ▪ Low activity ▪ Very low activity ▪
11. Smoking (cig)	▪ *Never smoked* ▪ Ex-smoker ▪ Light ▪ Moderate ▪ Heavy
12. Educational level (hedqul08)	▪ *Degree or higher* ▪ HNC/D ▪ Higher (higher grade or equivalent) ▪ School grade (standard grade or equivalent) ▪ No qualifications

13. Social class (schrpg7)	▪ *Professional* ▪ Managerial technical ▪ Skilled non-manual ▪ Skilled manual ▪ Semi-skilled manual ▪ Unskilled manual-Others
14. SIMD quintile (simd20_rp)	▪ *Least deprived* ▪ 4th ▪ 3rd ▪ 2nd ▪ Most deprived

For rural cases, the survey years 2018, 2019, and 2021, similar to trends observed in the urban & rural and urban datasets, show increased odds of the outcome compared to 2017 (Figure 3.17). The odds ratios (ORs) are 1.356 (CI: 1.008–1.825), 1.470 (CI: 1.099–1.967), and 2.015 (CI: 1.508–2.693), respectively (Figure 3.17). Gender also demonstrates significant differences, with females having significantly lower odds of the outcome than males (OR: 0.473, CI: 0.390–0.574). Age is a recurring determinant, with the highest odds observed in the 45–54 age group (OR: 2.348, CI: 1.644–3.356), followed by the 55–64 (OR: 2.260, CI: 1.563–3.269) and 35–44 (OR: 1.704, CI: 1.178–2.465) groups, when compared to the 25–34 reference group.

Birthplace also plays a significant role, as individuals born in the 'Rest of the UK' (OR: 0.430, CI: 0.274–0.675) and 'Elsewhere' (OR: 0.145, CI: 0.063–0.332) exhibit lower odds of the outcome compared to those born in Scotland. Unexpectedly, a contradictory result is observed for the Ethnicity variable, where individuals identified as 'White: Rest UK' have higher odds (OR: 1.812, CI: 1.145–2.866) than 'White: Scotland'. It is also noticeable that the 'Other ethnicities' category has wide confidence intervals (OR: 0.898, CI: 0.408–1.977), indicating substantial uncertainty and potentially low statistical power for this group. Religious affiliation also impacts the outcome, as individuals associated with the 'Church of Scotland' (OR: 0.544, CI: 0.422–0.702), 'Roman Catholic' (OR: 0.610, CI: 0.403–0.924), and 'Other religions' (OR: 0.465, CI: 0.329–0.658) have reduced odds compared to those with 'No religion'. Marital status shows that separated or widowed individuals have significantly lower odds of the outcome (OR: 0.390, CI: 0.265–0.573) compared to those who are married.

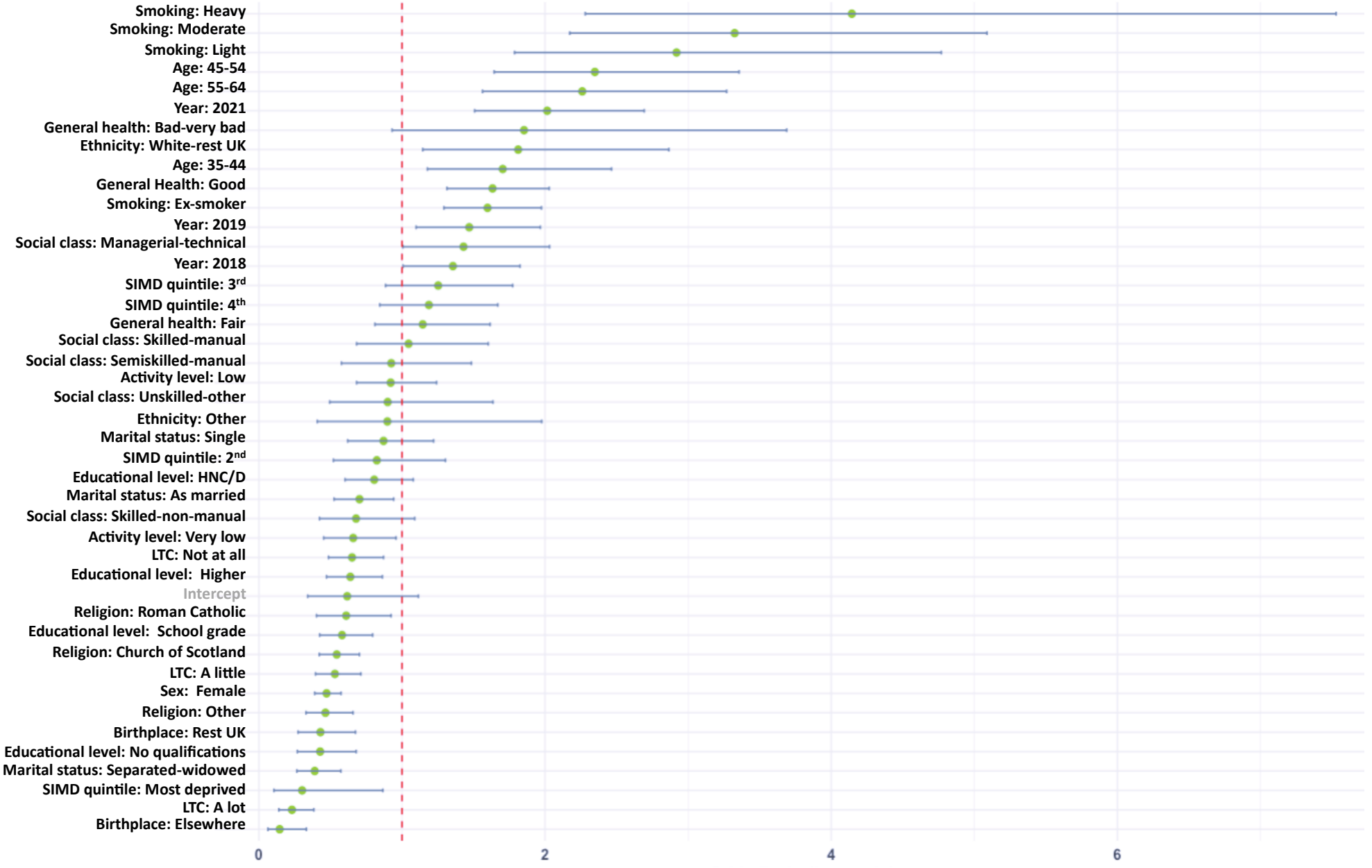


FIGURE 3.17 RURAL MODEL: ODDS RATIOS AND 95% CONFIDENCE INTERVALS

Self-assessed health status ('General Health') indicates that individuals reporting 'Bad–Very bad' health have the highest odds (OR: 1.853, CI: 0.931–3.689) of all categories compared to those reporting 'Very good' health. Long-term conditions ('LTC') are associated with decreasing odds of the outcome as limitations become more severe. Individuals with 'A lot of limitation' are the least likely to experience the outcome (OR: 0.231, CI: 0.139–0.384) compared to those without limitations. Activity levels indicate that individuals with 'Very low' activity have reduced odds (OR: 0.659, CI: 0.453–0.959) compared to those meeting physical activity recommendations. Smoking status significantly influences the outcome, with higher odds observed for 'Ex-smokers' (OR: 1.598, CI: 1.293–1.975), 'Light smokers' (OR: 2.919, CI: 1.787–4.769), 'Moderate smokers' (OR: 3.325, CI: 2.172–5.089), and 'Heavy smokers' (OR: 4.144, CI: 2.281–7.529) compared to 'Never smokers'. The considerably high odds for 'Heavy smokers' highlight the strong and escalating association between smoking intensity and the outcome.

Educational level is also a key determinant. Individuals with 'Higher grade' (OR: 0.639, CI: 0.473–0.863), 'Standard grades' (OR: 0.581, CI: 0.425–0.795), or 'No qualifications' (OR: 0.428, CI: 0.269–0.680) have lower odds of the outcome compared to those with 'Degree or HNC/D'.

Deprivation, as assessed by SIMD quintiles, reveals that individuals in the most deprived quintile have significantly lower odds (OR: 0.302, CI: 0.105–0.867) compared to those in the least deprived quintile. However, the 4th quintile (OR: 1.188, CI: 0.845–1.670) and 3rd quintile (OR: 1.253, CI: 0.885–1.774) have odds closer to the least deprived group.

3.10.4 EVALUATING THE RURAL MODEL IN THE TEST SUBSET

The ROC value (65.10%) in the rural dataset is the lowest obtained from the three models (Table 3.16), indicating a very low ability to distinguish between those who consume more than 14 units and those who consume 14 or fewer units per week. The sensitivity (74.52%) is the highest among all the metrics for the rural datasets, similar to the models already analysed, but it is imbalanced with the specificity, which is only 47.57%. This imbalance results in a considerable number of false positives. The model does not stand out for its accuracy either, with 54.07% of the total predictions being correct. The PPV is the most affected, with a value of only 31.12% because of the false positive, indicating that the model is

TABLE 3.16 RURAL DATASET: METRICS IN THE TEST SUBSET

ROC	Sensitivity	Specificity	Accuracy	Precision	F1-score
65.10%	74.52%	47.57%	54.07%	31.12%	43.90%

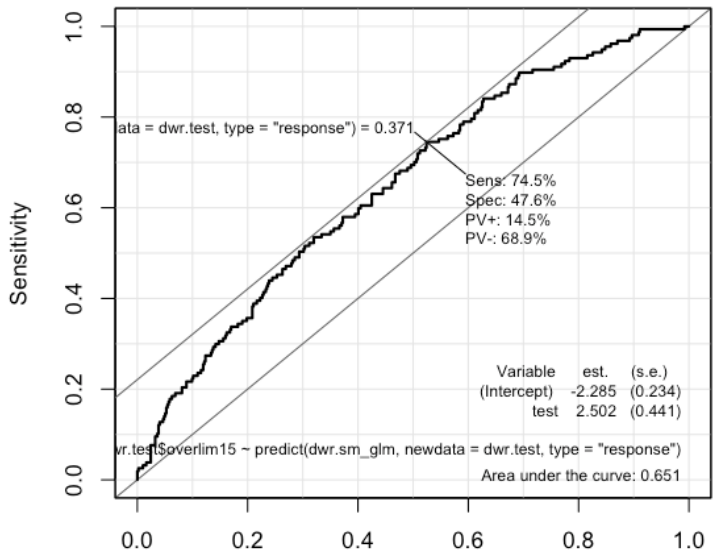


FIGURE 3.18 RURAL TEST SUBSET: MODEL'S EVALUATION RESULTS

often incorrect when predicting a positive case. The F1-score (43.90%) is below 50%, reflecting the balance between poor precision and recall (Figure 3.18).

3.11 CONSIDERATIONS ABOUT THE THREE MODELS

This section presents the findings from three distinct models: the urban & rural model, the urban-only model, and the rural-only model. It explores the key variables influencing alcohol consumption patterns, including demographic factors, health status, and socioeconomic indicators. The results and interpretation of the LR models are discussed, with a focus on predictors of harmful alcohol consumption exceeding the recommended weekly limit of 14 units. The performance of these models in the test subset is also evaluated.

3.11.1 VARIABLES CONSIDERED IN THE THREE MODELS

The model comprising both urban & rural cases contained 15 variables, while the urban and rural-only models each included 14 variables. The following variables were common across all three final models:

- Survey year (syear)
- Age (ag16g10)
- Sex (sex)
- Ethnicity (ethnic05)
- Religion (religi04)
- Marital status (maritalg)
- General health (genhelf)
- LTC (limitac_h)
- Activity level (adt10gptw)
- Smoking (cig)
- Educational level (hedqul08)
- Social class (schrpg7)
- SIMD quintile (simd20_rp)

The models with the best performance, selected from datasets containing only urban or only rural cases, were identical in terms of their variable composition. However, the model based on the entire population (without distinguishing between urban and rural areas) included the variable Economic Activity (neconacb), which reflects employment status, and lacked the variable Birthplace (birthpla3), present in the other two models.

3.11.2 RESULT INTERPRETATION IN THE THREE MODELS

In all three settings, the odds of harmful alcohol consumption increased over the years (syear) compared to 2017, particularly in 2021. The odds ratios (ORs) were 2.142 (CI: 1.887–2.433) in the urban & rural model, 1.875 (CI: 1.630–2.156) for urban, and 2.015 (CI: 1.508–2.693) for rural, suggesting a temporal trend potentially influenced by COVID-19 (see Section 4.2).

Gender (sex) disparities were evident across the models, with females consistently showing lower odds than males. The ORs for females were 0.456 (CI: 0.420–0.496) in the urban & rural model, 0.424 (CI: 0.386–0.465)

for urban, and 0.473 (CI: 0.390–0.574) for rural, suggesting that, in all settings, females were less likely to engage in harmful alcohol consumption (more than 14 units per week) compared to males.

Age remained a critical determinant, with the highest odds observed in the 45–54 and 55–64 age groups. Across all models, these age groups showed ORs above 2 compared to the reference 25–34 group, highlighting those older individuals were notably more likely to report exceeding the recommended alcohol limits.

Ethnicity (ethnic05) and religion (religi04) also played significant roles, although the specific patterns differed between settings. For instance, individuals from the 'White: Rest of UK' ethnic group had a higher probability of exceeding the 14-week limit than those from the 'White from Scotland' reference category in rural areas (OR: 1.812, CI: 1.145–2.866), but showed lower odds in the urban & rural model (OR: 0.720, CI: 0.636–0.816) and no significant difference in the urban model (OR: 1.160, CI: 0.910–1.479). The ethnicity group labelled 'Other' (not from Scotland or the UK) exhibited a very low OR in the urban & rural model (OR: 0.376, CI: 0.315–0.449) and the urban model. However, in the rural model, the OR approached 1 (0.898), though the 95% CI was wide (0.408–1.977), indicating potential issues with the estimate, possibly due to a small sample size or insufficient representation of this ethnic group in the rural setting, leading to greater variability in the OR.

Regarding religious denominations, the 'Other religions' group showed consistently lower odds than those who declared not professing any religion, with values below 0.5. The ORs for 'Other religions' were 0.399 (CI: 0.335–0.474) in the urban & rural combined model, 0.477 (CI: 0.393–0.578) in the urban model, and 0.465 (CI: 0.329–0.658) in the rural model.

In the rural model, marital status (maritalg) showed that individuals identified as separated, widowed, or divorced had a very low OR (OR: 0.390, CI: 0.265–0.573) compared to the married reference group. However, in the urban & rural model (OR: 0.724, CI: 0.627–0.836) and the urban model (OR: 0.841, CI: 0.719–0.984), this difference was less pronounced, with ORs closer to 1, indicating a minor effect of marital status in these settings in comparison with the rural area.

Self-assessed health status (genhelf) showed varying associations with harmful alcohol consumption across the three models. In the urban & rural combined model, individuals reporting 'Bad' or 'Very bad' health had higher odds (OR: 1.316, CI: 1.025–1.690). However, in the urban model, the association was weak (OR: 1.089, CI: 0.840–1.413), while in the rural model, the odds were significantly higher (OR: 1.853, CI: 0.931–3.689). These results suggested that poor self-reported health was more strongly associated with harmful alcohol consumption in rural areas.

Very low physical activity levels (adt10gptw) were associated with lower odds compared to those meeting physical activity recommendations across the three settings (OR: 0.616, CI: 0.539–0.703 for urban & rural, OR: 0.576, CI: 0.490–0.676 for urban, and OR: 0.659, CI: 0.453–0.959 for rural). Also, severe limitations (limitac_h) significantly decreased the odds of harmful alcohol consumption across all contexts. The most pronounced effect was observed in the rural model (OR: 0.231, CI: 0.139–0.384), indicating that individuals with serious long-term conditions (LTC), especially in rural areas, were less likely to exceed the alcohol consumption limits.

Smoking behaviour (cig) had a significant impact on harmful alcohol consumption, with ex-smokers, light smokers, moderate smokers, and heavy smokers all showing higher odds compared to non-smokers. The ORs for heavy smokers were the highest, with 2.558 (CI: 2.035–3.215) in the urban & rural model, 2.748 (CI: 2.150–3.512) in the urban model, and 4.144 (CI: 2.281–7.529) in the rural model, indicating a particularly strong association between heavy smoking and harmful alcohol consumption, particularly in rural areas.

Educational attainment (hedqul08) consistently showed that lower qualifications were associated with lower odds of harmful alcohol consumption compared to higher education levels, especially in the rural model, where individuals with no qualifications showed the lowest OR of all categories in the three settings (OR: 0.428, CI: 0.269–0.680). Social class categories (schrpg7) generally had ORs close to 1, showing no clear difference from the professional reference category, except for the managerial-technical category in the three models (OR: 1.479, CI: 1.279–1.709 in the urban & rural model; OR: 1.293, CI: 1.096–1.525 in the urban model; and OR: 1.430, CI: 1.007–2.031 in the rural model). Additionally, only in the rural setting did the 'Skilled non-manual' category show a lower OR than the professional reference category (OR: 0.679, CI: 0.424–1.089).

Those in the most deprived areas (simd20_rpa) showed the lowest ORs in the urban (OR: 0.585, CI: 0.499–0.687) and rural (OR: 0.302, CI: 0.105–0.867) models. In the urban & rural model, the second most deprived quintile showed the lowest OR (OR: 0.590, CI: 0.516–0.675), followed closely by the most deprived quintile (OR: 0.639, CI: 0.550–0.743).

Overall, the three models showed similar ORs for self-reported alcohol consumption. The odds of exceeding the 14-week limit increased over the years, with the highest OR observed in 2021, reflecting an upward trend in harmful alcohol consumption. Age, smoking status, educational attainment, and social class were all significant predictors of harmful alcohol consumption, with older individuals, smokers, and those with higher educational attainment showing higher odds of exceeding the limit.

In contrast, females, individuals with severe long-term conditions, those living in the most deprived areas, those who professed a religion (particularly 'Other religions'), those with very low physical activity levels, and those who were separated, widowed, or divorced were less likely to report exceeding the 14-week alcohol consumption limit.

3.11.3 EVALUATING THE THREE MODELS IN THE TEST SUBSETS

The models derived from oversampled and SMOTE datasets generally demonstrated superior performance compared to those built on the original data.

The area under the receiver operating characteristic curve (ROC) for all models remained below the generally acceptable threshold of 70% for predictive classification models. The ROC was lowest for the rural-only model (65.10%) and highest for the urban model (66.90%) (Tables 3.8, 3.12, and 3.16). Sensitivity was relatively high for the urban (73.17%) and rural (74.52%) models but lower for the urban & rural model (66.93%). Specificity was generally low across all models: urban & rural combined (56.88%), urban (59.29%), and rural (47.57%). Accuracy did not exceed 60% in any of the models, with urban & rural showing the highest accuracy (59.29%),

followed by urban (57.35%) and rural (54.07%). Positive Predictive Value (PPV) was similarly low, hovering around 30% across all three models (urban & rural = 32.85%, urban = 32.35%, and rural = 31.12%). The poor performance reflected in the F1-score, which did not exceed 45% in any model, was indicative of the challenges faced by these models in effectively predicting harmful alcohol consumption (urban & rural = 44.07%, urban = 44.87%, and rural = 43.90%).

3.12 RF IN THE RURAL DATASETS

An RF was also considered to compare the results. Table 3.17 summarises the performance of the RF model across the rural training datasets to find the optimal hyperparameters.

TABLE 3.17 RURAL DATASETS: RF RESULTS

Dataset	N trees	mtry	Node Size	Sample Size	OOB_RMSE
dwr.train	200	2	5	0.5	0.487
up30_dwr	500	4	3	1	0.334
up40_dwr	500	4	3	1	0.310
up50_dwr	500	6	3	1	0.285
smk3r50_dwr	500	5	3	1	0.410
smk3r75_dwr	500	4	3	1	0.344
smk3r100_dwr	200	5	5	1	0.335
smk5r50_dwr	500	4	3	1	0.409
smk5r75_dwr	500	6	3	1	0.368
smk5r100_dwr	300	7	3	1	0.353
smk10r50_dwr	200	3	3	1	0.411
smk10r75_dwr	500	7	3	1	0.362
smk10r100_dwr	500	6	5	1	0.352

Most datasets showed improved performance with an increased number of trees (n_trees = 500), which reduced variance and helped avoid overfitting, a key benefit of ensemble methods like random forests. The 'mtry' values, which define the number of features considered at each split, ranged from two to seven, indicating that the optimal number of features varied by dataset. In most cases, the node size remained consistent at 3, suggesting that smaller node sizes generally led to better performance. However, a few datasets, such as ‘dwr.train’, ‘smk3r100_dwr’ and ‘smk3r100_dwr’ showed improved performance with a larger node size (5), which may be attributed to the nature of the data and the balance between bias and variance when partitioning the data into nodes.

In terms of sample size, the overall model performance was best when using the full dataset (Sample Size = 1), except for the original dataset (‘dwr.train’), where a smaller sample size might have been more beneficial due to potential class imbalance or data quality issues. Oversampling techniques (like SMOTE) produced some of the best performance metrics, with the lowest OOB_RMSE values observed in the oversampled datasets ‘up30_dwr’ (0.334), ‘up40_dwr’ (0.310), and ‘up50_dwr’ (0.285). These datasets outperformed the baseline by using a consistent set of parameters (n_trees = 500, mtry = 4 or 6, node size = 3, sample size = 1), suggesting that oversampling improved the model's ability to generalise to minority class observations.

However, datasets generated via SMOTE showed higher OOB_RMSE values, with 'smk3r50_dwr' (0.410), 'smk5r50_dwr' (0.409), and 'smk10r50_dwr' (0.411), indicating that synthetic samples created with SMOTE at or = 0.5 did not significantly reduce the model's error rate. This could imply that the SMOTE technique, at this particular ratio, failed to generate sufficiently realistic or diverse samples that improved model accuracy, possibly due to overfitting or poor quality of synthetic samples.

The model from the oversampled dataset 'up50_dwr', which utilised a 50:50 ratio between the majority and minority classes, achieved the lowest OOB_RMSE of 0.285, indicating superior generalisation capability and predictive accuracy. The performance of this model demonstrated that balancing the class distribution through oversampling was effective in improving the random forest's ability to classify minority class instances more accurately, without introducing additional bias.

Cross-validation (k = 10, repetitions = 5) was performed using the best hyperparameters previously selected (number of trees = 500, number of features = 6, node size = 3, sample size = 1). The results of cross-validation on the training subsets are presented in Table 3.18, showing consistently strong performance. The model demonstrated excellent predictive performance, with ROC (96.9%, SD = 0.013) and AUC (95.0%, SD = 0.020) above 90%.

TABLE 3.18 RF: CROSSVALIDATION RESULT

Metrics	Value	SD
ROC	96.9%	0.013
AUC	95.0%	0.020
Sensitivity	86.7%	0.026
Specificity	94.9%	0.024
Accuracy	90.8%	0.018
PPV	94.5%	0.024
Recall	86.7%	0.026
F1-score	90.4%	0.019

Sensitivity (86.7%, SD = 0.026) and specificity (94.9%, SD = 0.024) were both high, indicating that the model was adept at correctly identifying both positive and negative cases. Accuracy was also strong at 90.8% (SD = 0.018), while the PPV was 94.5% (SD = 0.024), reflecting the model's ability to correctly identify positive instances. The recall (86.7%, SD = 0.026) and F1-score (90.4%, SD = 0.019) further confirmed the model's robustness in maintaining a good balance between sensitivity and precision.

These results suggested that the oversampling approach, in combination with the random forest algorithm, provided a highly effective model for classification tasks in imbalanced datasets, particularly in the rural context.

The ranking of importance (Figure 3.19) obtained through the model showed that the six more important variables are sex, cig (smoking habits), limitac_h (LTC), ag16ag10 (age), hedqul08 (educational level), schrpg7 (social class classification).

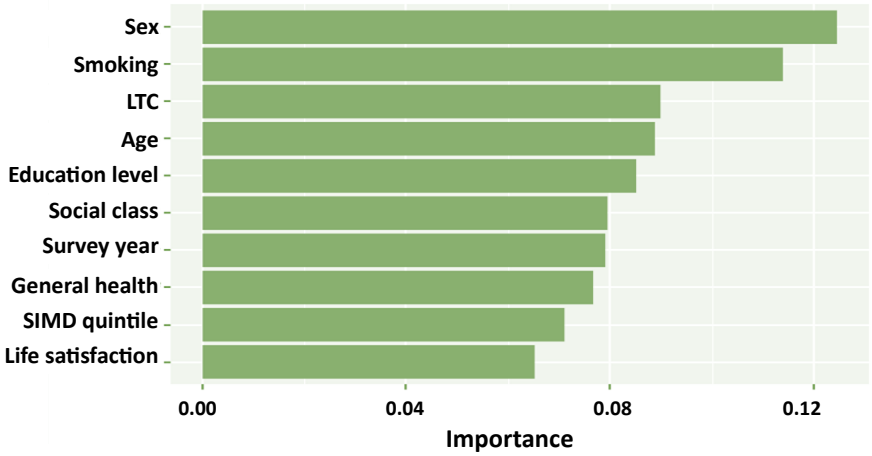


FIGURE 3.19 VARIABLE IMPORTANCE FROM RF

However, the model's performance on the test subset showed a significant decline compared to its performance during training (Table 3.19). The sensitivity (recall) dropped to only 17.83%, indicating that the model had a very low ability to correctly identify positive instances, resulting in a high false negative rate (Type II error). While the specificity remained relatively high at 85.63%, suggesting that the model was effective at identifying negative cases, the overall accuracy was 69.28%. This accuracy was likely influenced by the model's strong performance in predicting the majority class correctly, rather than its ability to classify positive cases.

TABLE 3.19 RURAL TEST SUBSET: RF METRICS

Metric	Value
Sensitivity	0.18
Specificity	0.86
Accuracy	0.69
Precision	0.28
F1 Score	0.22

The positive predictive value (PPV) was notably low at 28.28%, indicating a high false positive rate and a large number of incorrect positive predictions. The F1-score, which combines both precision and recall, was also low at 21.87%, reflecting a poor balance between PPV and recall. This suggests that the model struggled significantly with predicting the positive class, highlighting a clear need for improvement, particularly in reducing false negatives and enhancing the prediction of positive instances.

4. DISCUSSION

This section discusses findings from the statistical analysis conducted across three datasets with cases from the urban and rural, only urban, and only rural areas. The goal is to understand how these factors interplay in different geographic contexts, considering the results obtained in this study. First, the discussion will be regarding the association between sociodemographic, health-related, educational, and socioeconomic variables and alcohol consumption, particularly about the 14-unit weekly limit, which is considered the threshold above which alcohol consumption is deemed harmful. Second, it will discuss the findings concerning differences between rural and urban areas. Finally, it will examine the results regarding the application of predicting modelling through the three datasets.

4.1 ALCOHOL CONSUMPTION PATTERNS CONCERNING SOCIODEMOGRAPHICS AND SOCIOECONOMICS FACTORS

The 25-34 age group had more representation in the group that reports consuming below the limit, with similar rates in the urban & rural, urban, and rural datasets. This observation is consistent with the research of Arnett (1998), which shows that individuals in their late twenties often undergo family role transitions, such as getting married or becoming parents. These transitions have been linked to reductions in risky behaviours, with married individuals (Curran et al., 1998) and parents showing lower levels of substance use and also lower risky driving and unsafe sexual practices. Conversely, in the above-limit consumption group, the datasets exhibit slight differences, with the 55-64 age group being predominant in urban areas (28.90%), while the 45-54 age group is more prominent in rural areas (26.75%). This disparity can be attributed to a relaxation of family responsibilities and an increase in economic status among these age groups (Smith and Foxcroft, 2009), which may enable them to afford and consume more alcohol.

The group consuming alcohol below the weekly limit exhibits a higher proportion of females across all datasets. In the urban & rural dataset, males have approximately 120% higher odds of exceeding the weekly alcohol limit than females, 134% in the urban and 108% in the rural. The observation aligns with previous international studies, indicating that globally, men are more likely to consume alcohol and tend to drink more heavily compared to women (McCaul et al., 2019, Grant et al., 2015). This trend can be attributed to shifts in women's roles and norms. Although the gender disparity in alcohol consumption has been changing in recent years, connected to social transformations that include increased educational attainment, higher employment rates outside the home, and delayed age at first marriage and childbearing (McCaul et al., 2019).

There are significant differences in consumption patterns among different birthplace, ethnic, and religious groups. Individuals born outside the UK, particularly those of Asian ethnicity and those who follow religions other than Roman Catholicism or the Church of Scotland, tend to report lower alcohol consumption. These findings are interconnected, as people born in certain places often belong to similar ethnic groups and may share the same religion. This highlights the importance of considering multiple demographic factors together, as they could collectively influence behaviours and attitudes towards alcohol use (Caetano, 1998, Caetano et al.,

1998). However, this finding is more relevant in the urban Scottish contexts, where these minorities are more represented than in the rural dataset.

Individuals who self-classify their general health as very poor, who have an LTC, who report life satisfaction below the mode, and whose level of physical activity is classified as very low, reported lower alcohol consumption relative to the rest of the categories. This finding may be connected to the fact that alcohol consumption is frequently associated with social interactions in various social contexts (Ritchie, 2007, Abbey et al., 1993, Smith and Foxcroft, 2009). These individuals with poor health are likely to have less social interaction than the rest of the groups, which may contribute to their lower alcohol consumption. Additionally, these individuals might avoid alcohol due to potential adverse effects on their existing health conditions, further contributing to the observed trend.

Non-smokers are inclined to consume less alcohol in the three datasets. Earlier studies corroborate this finding (Bhunu and Mushayabasa, 2012, Romberger and Grant, 2004, Shiue, 2014). Alcoholism is estimated to be ten times more common among smokers than non-smokers. Cigarette smoking is prevalent among individuals with alcohol dependence, with up to 80% of those affected also being smokers. This prevalence is approximately two to three times higher than in the general population. This combination of risky behaviours can lead to severe health issues, including a heightened risk of diseases such as head and neck cancers (Romberger and Grant, 2004).

The findings indicate a higher rate of harmful alcohol consumption among individuals with higher education levels and those in the least deprived areas in the three datasets. This trend aligns with previous investigations (Christensen et al., 2017, van Oers et al., 1999, Grittner et al., 2013, Zhou et al., 2021, Beard et al., 2019), which suggest that individuals of higher socioeconomic status (SES) tend to drink smaller amounts more frequently. A Mendelian randomisation study using UK Biobank data (Zhou et al., 2021) found that one additional year of education is linked to increased total alcohol intake and frequent drinking, especially wine and champagne, and often with meals. Regardless, the researchers also found that despite higher alcohol intake, more educated individuals tend to adopt healthier drinking practices. However, this should be taken cautiously. In an international comparison of social inequalities in alcohol use across countries with different levels of economic development, Grittner et al. (2013) found that, in the majority of the countries studied, those with higher education were more likely to be current drinkers. These patterns may be attributed to the greater economic capacity of higher socio-economic status (SES) individuals to afford more frequent alcohol consumption. This frequent consumption, even in lower quantities, can potentially exceed the weekly limit of 14 units, which underscores the need for targeted interventions to reduce alcohol intake among highly educated populations.

However, it should be underlined that globally, the relationship between SES and alcohol consumption shows inconsistencies (van Oers et al., 1999). Discrepancies may stem from varying research contexts across different countries and regions, emphasising the influence of diverse drinking cultures and attitudes towards alcohol-

related issues. Methodological differences, especially in defining heavy alcohol consumption, also play a significant role in interpreting results. A systematic investigation into this theme could provide valuable insights.

Additionally, it should be noted that several studies have found that very high consumption and alcohol-related medical events were more pronounced in the lower-educated group (Christensen et al., 2017, van Oers et al., 1999). Although this study did not analyse very high consumption, future research could explore this area, considering both urban and rural contexts, because it is precisely these hazardous drinkers who experience the worst complications related to alcohol consumption, including increased morbidity, hospitalisations, and mortality.

Wine is the predominant beverage in both the above-limit groups and below-limit groups across all datasets, with particularly high consumption in the rural dataset. Ritchie (2007) highlights that since wine sales began through supermarkets in the UK during the 1970s, wine consumption has more than doubled, making the UK the largest wine import market by value globally. Currently, around 61% of UK adults regularly consume wine. Ritchie (2007) notes that wine is frequently enjoyed in various social contexts, such as restaurants, dinner parties, and as gifts. The social aspect of wine consumption is significant, serving as a rite of passage and fostering social interaction (Ritchie, 2007, Smith and Foxcroft, 2009). This trend of increasing wine consumption is not limited to the UK. About 80% of countries are showing similar trends in wine imports, a clear sign of the impact of globalisation on the wine market and the growing demand for foreign wine (Ohana-Levi and Netzer, 2023). This tendency seems particularly strong in modern, developed, and wealthy countries, characterised by urban populations and high incomes. Therefore, the major forces driving the wine market are economic growth and wider competition.

4.2 PREDICTING MODELLING ACROSS THE THREE DATASETS

The analysis identified several consistent predictors across all models. Variables such as sex, age (ag16g10), ethnicity (ethnic05), religion (religi04), health limitations (limitac_h), educational levels (hedqul08), socioeconomic status (simd20_rpa), and smoking habits (cig) were significant in predicting the outcome. These variables reflect the multifaceted nature of the factors influencing the outcome, encompassing demographic, socioeconomic, and lifestyle aspects.

The models have also brought to light variations in the significance of predictors depending on whether the analysis included all cases together or was stratified by urban and rural areas. In the model based on the combined dataset, the variable 'economic activity', which identifies whether an individual is employed or unemployed, emerged as a significant predictor. However, this variable was not included in the models generated for the urban and rural datasets. Conversely, the variable 'birthplace', associated with the individual's place of birth, was included in the urban and rural models but was not a significant predictor in the combined dataset. However, the predictors identified for the only-urban and only-rural datasets were consistent with each other. These findings underscore the importance of developing tailored predictive models for distinct populations, as characteristics and predictor significance can vary based on the specific context.

Additionally, some categories were poorly represented in certain datasets, particularly in the rural dataset, which had a smaller sample size and less diversity in certain demographic and socioeconomic groups. This limited representation may have constrained the model's ability to generalise patterns effectively, highlighting the need for caution when interpreting the results.

The years 2018, 2019, and 2021 were associated with higher odds of the outcome than the reference year 2017. This trend could be related to the impact of COVID-19, particularly in 2021, where the pandemic may have exacerbated certain behaviours or stressors contributing to higher alcohol consumption (Althobaiti et al., 2021). Additionally, the SHeS had to adjust their fieldwork to meet pandemic requirements, which could have influenced the data collection and subsequent findings (ScotCen Social Research, 2021c). Further research should investigate this temporal trend and consider specific survey application factors that may influence these results.

4.3. LIMITATIONS

The data imbalance adversely affected the results of the predictive modelling. Using oversampling and SMOTE techniques improved model performance, particularly in urban datasets. However, the models faced challenges balancing sensitivity and specificity, resulting in moderate overall performance. The evaluation of the LR models indicated moderate discrimination ability, with AUC ROC values hovering around 70%. The models demonstrated high sensitivity but struggled with specificity, leading to a higher rate of false positives. This imbalance suggests that while the models effectively identify positive cases, they incorrectly classify many negative cases as positive.

The poor performance of the predictive models can be attributed not only to imbalanced data but also to the inherent limitations of self-reported alcohol consumption data collected through surveys (Nugawela et al., 2016, Boniface and Shelton, 2013, Gilligan et al., 2019). Despite the importance of understanding alcohol consumption within a population to assess the effectiveness of public policies and the utility of national surveys in collecting comprehensive data on alcohol use alongside sociodemographic, socioeconomic, and health information, these surveys present significant challenges. Studies estimate that self-reported alcohol consumption underestimates actual sales by approximately 40 to 60%. The SHeS reports acknowledge this bias, clarifying that the survey's estimate of alcohol consumption has consistently been below the data collected from sales or tax revenues related to alcohol (2018a, ScotCen Social Research, 2017b, 2019a, 2021b).

The Scottish Health Survey (SHeS) has recorded a decline in the self-reported mean number of units of alcohol consumed by adults per week (ScotCen Social Research, 2022a). In 2003, the average consumption was 16.1 units, whereas in 2022, the most recent published result, it had decreased to 12.6 units. However, alcohol-related deaths have shown an overall increase in Scotland since 2012 (NRS, 2023). In 2022, 1,276 deaths were attributable to alcohol-specific causes, representing a 2% increase from 2021 and the highest annual death toll since 2008. Additionally, despite 2020 and 2021 witnessing the lowest recorded levels of alcohol sales in Scotland since 1994, each adult still purchased an average of 9.4 litres of pure alcohol, amounting to 18.1 units

of alcohol per adult per week. This consumption rate exceeds the UK Chief Medical Officers' guideline of 14 units per adult per week by nearly 30%.

Furthermore, it is challenging to convert respondents' consumption volumes to the alcohol units. The SHeS reports highlight this difficulty, noting the various obstacles in calculating units at a population level (ScotCen Social Research, 2022b). Alcohol concentration can vary greatly among different types of beverages and over time (ScotCen Social Research, 2022b, Gilligan et al., 2019). For instance, wine can have an alcohol content ranging from 8% to 13% v/v, while spirits can range from 37.5% v/v to 57.3% v/v. (Gilligan et al., 2019). Moreover, accurately estimating the volume of a consumed drink is also challenging (Gilligan et al., 2019, Devos-Comby and Lange, 2008), and the survey reports address this by clarifying that the reported volumes have not been validated (ScotCen Social Research, 2022b). For example, a standard measure of wine in a pub is typically 175 ml, but it can be double that amount in a restaurant if larger glasses are used (Gilligan et al., 2019). Drinkers may also underestimate their actual consumption (Devos-Comby and Lange, 2008). These differences make subjective assessments of alcohol content very challenging. To all of the above, it must be added other factors, such as nonresponse bias and sampling frame limitations also contribute to the underestimation (Nugawela et al., 2016). These issues highlight the need for careful interpretation and adjustment when using survey data to inform public health strategies.

5. CONCLUSION

In conclusion, most people represented in this study consume alcohol below the 14-unit weekly limit, showing similar patterns across the urban and rural datasets. Individuals more likely to consume above the limit include older age groups, males, those in good health, those without long-term conditions (LTCs), those with higher life satisfaction, smokers, individuals with higher education, and those in the least deprived areas. Wine is the predominant beverage for both below- and above-limit consumption groups. The differences between urban and rural populations are not substantial enough to indicate significant discrepancies in alcohol consumption behaviour.

The low positive predictive value (PPV) and F1-score for the logistic regression (LR) model further underscore the challenges in achieving reliable predictions. The performance of the models suggests that while LR is a valuable tool, its effectiveness may be constrained by the inherent complexities and imbalances in the dataset.

The significant disparity between the cross-validation results (Table 3.18) and the test subset metrics (Table 3.19) for the random forest (RF) model in the rural dataset suggests potential overfitting to the training data, leading to poor generalisation of unseen data. This discrepancy may also arise from data imbalance within the test subset, differing class distributions, or feature variability between the rural training and test subsets. Cross-validation averages metrics across multiple folds, reducing the impact of anomalies, whereas the test subset, being a single partition, may not fully represent the underlying data distribution. Additionally, the relatively small dataset size of the rural test subset (665 cases) can exacerbate these issues, limiting the model's ability to accurately capture and generalise patterns, further contributing to the observed performance differences.

In future research, it would be beneficial to include additional variables excluded in this study due to extensive missing data. In particular, the Alcohol Use Disorders Identification Test (AUDIT) score (Babor, 2001) could provide valuable insights into alcohol consumption patterns and the risk of alcohol-related harm. The AUDIT score is a widely used screening tool for identifying individuals with harmful alcohol use, based on a series of 10 questions regarding alcohol consumption, dependence symptoms, and alcohol-related problems. Integrating AUDIT scores into predictive models could enhance their ability to identify at-risk individuals more accurately.

For handling missing data, the application of the Multiple Imputation by Chained Equations (MICE) package (Buuren and Groothuis-Oudshoorn, 2011) could be considered. While an initial attempt was made during this study, exploratory results did not yield improved metrics. Nevertheless, revisiting this approach could enhance the model's performance and provide more comprehensive insights. Moreover, multiple imputation is an alternative to listwise deletion and may provide a more robust solution to handling missing values, which could reduce the potential bias introduced by data loss.

Also, although stepwise selection was chosen in this study due to its ease of use and interpretability, alternative approaches such as LASSO or elastic net should be considered in future analyses (Ranstam and Cook, 2018). These methods may offer advantages in handling multicollinearity and improving model stability, particularly for datasets with numerous predictors. Evaluating the performance of models generated using these techniques,

alongside metrics such as AUC and cross-validation, could further enhance the robustness and generalisability of the results. Adopting these approaches would provide additional transparency and rigour to the variable selection process while addressing some of the inherent limitations of stepwise selection.

In conclusion, the identified significant predictors and their odds ratios (OR) provide valuable insights into the factors influencing the outcome, underscoring the need for tailored predictive modelling approaches. Given the observed discrepancies and limitations, future research should prioritise refining these models and investigating advanced modelling techniques to improve predictive accuracy and reliability. Additionally, exploring alternative approaches will enhance the applicability of these models across diverse populations, ensuring more robust and generalisable findings.

REFERENCES

ABBEY, A., SMITH, M. J. & SCOTT, R. O. 1993. The relationship between reasons for drinking alcohol and alcohol consumption: an interactional approach. *Addict Behav,* 18**,** 659-70.

AKAIKE, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* 19**,** 716-723.

ALCOHOL FOCUS SCOTLAND (AFS) 2021. Tackling harm from alcohol. Alcohol policy priorities for the next Parliament.: AFS,.

ALCOHOL FOCUS SCOTLAND (AFS) 2023. Response to Health, Social Care and Sport Committee consultation on healthcare in remote and rural areas.: AFS,.

ALCOHOL HEALTH ALLIANCE (AHA) 2023. Pouring over Public Opinion: Alcohol Policies in the UK. AHA,.

ALLISON, P. 2002. Missing Data. Thousand Oaks, California.: SAGE Publications, Inc.

ALTHOBAITI, Y. S., ALZAHRANI, M. A., ALSHARIF, N. A., ALROBAIE, N. S., ALSAAB, H. O. & UDDIN, M. N. 2021. The Possible Relationship between the Abuse of Tobacco, Opioid, or Alcohol with COVID-19. *Healthcare,* 9**,** 2.

ARNETT, J. J. 1998. Risk behavior and family role transitions during the twenties. *Journal of Youth and Adolescence,* 27**,** 301-320.

BABOR, T. F. H.-B. J. C. S. J. B. M. M. G. 2001. AUDIT - The alcohol use disorders identification test: guidelines for use in primary care. Second Edition ed. Geneva: WHO,.

BEARD, E., BROWN, J., WEST, R., KANER, E., MEIER, P. & MICHIE, S. 2019. Associations between socio-economic factors and alcohol consumption: A population survey of adults in England. *PLoS One,* 14**,** e0209442.

BHATTACHARYA, A. 2023a. Getting in the spirit? Alcohol and the Scottish economy. London: The Social Market Foundation,.

BHATTACHARYA, A. 2023b. Minimum unit pricing and the messiness of evidence-based policy. *Addiction,* 118**,** 1617-1618.

BHUNU, C. P. & MUSHAYABASA, S. 2012. A Theoretical Analysis of Smoking and Alcoholism. *Journal of Mathematical Modelling and Algorithms,* 11**,** 387-408.

BIAU, G. & SCORNET, E. 2016. A random forest guided tour. *Test,* 25**,** 197-227.

BLOCKER, J. S., JR. 2006. Did prohibition really work? Alcohol prohibition as a public health innovation. *Am J Public Health,* 96**,** 233-43.

BONIFACE, S. & SHELTON, N. 2013. How is alcohol consumption affected if we account for under-reporting? A hypothetical scenario. *Eur J Public Health,* 23**,** 1076-81.

BORDERS, T. F. & BOOTH, B. M. 2007. Rural, suburban, and urban variations in alcohol consumption in the United States: findings from the National Epidemiologic Survey on Alcohol and Related Conditions. *J Rural Health,* 23**,** 314-21.

BOWER, J. 2016. Scotch Whisky: History, Heritage and the Stock Cycle. *Beverages,* 2**,** 11.

BRYDEN, A., ROBERTS, B., PETTICREW, M. & MCKEE, M. 2013. A systematic review of the influence of community level social factors on alcohol use. *Health Place,* 21**,** 70-85.

BURNS, N., PARR, H. & PHILO, C. 2002. Alcohol and mental health. *Findings paper* [Online]. Available: http://eprints.gla.ac.uk/96762/1/96762.pdf [Accessed 15/06/2024].

BUUREN, S. & GROOTHUIS-OUDSHOORN, K. 2011. mice: Multivariate Imputation by Chained Equations inR. *Journal of Statistical Software,* 45**,** 1 - 67.

CAETANO, R. 1998. Cultural and subgroup issues in measuring consumption. *Alcohol Clin Exp Res,* 22**,** 21S-28S.

CAETANO, R., CLARK, C. L. & TAM, T. 1998. Alcohol consumption among racial/ethnic minorities: theory and research. *Alcohol Health Res World,* 22**,** 233-41.

CARYL, F. M., PEARCE, J., MITCHELL, R. & SHORTT, N. K. 2022. Inequalities in children's exposure to alcohol outlets in Scotland: a GPS study. *BMC Public Health,* 22**,** 1749.

CHANDLER, A. & NUGENT, B. 2016. Alcohol Stories: a lifecourse perspective on self-harm, suicide and alcohol use among men. *In:* ALCOHOL RESEARCH UK (ed.). London.

CHATTERJEE, S. & HADI, A. S. 2006. Regression Analysis by Example. United States: John Wiley & Sons, Incorporated.

CHAWLA NITESH, V. 2002. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res,* 16**,** 321.

CHOWDHURY, M. Z. I. & TURIN, T. C. 2020. Variable selection strategies and its importance in clinical prediction modelling. *Fam Med Community Health,* 8**,** e000262.

CHRISTENSEN, H. N., DIDERICHSEN, F., HVIDTFELDT, U. A., LANGE, T., ANDERSEN, P. K., OSLER, M., PRESCOTT, E., TJONNELAND, A., ROD, N. H. & ANDERSEN, I. 2017. Joint Effect of Alcohol Consumption and Educational Level on Alcohol-related Medical Events: A Danish Register-based Cohort Study. *Epidemiology,* 28**,** 872-879.

COLLABORATORS, G. B. D. A. 2018. Alcohol use and burden for 195 countries and territories, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet,* 392**,** 1015-1035.

COURONNE, R., PROBST, P. & BOULESTEIX, A. L. 2018. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics,* 19**,** 270.

CURRAN, P. J., MUTHEN, B. O. & HARFORD, T. C. 1998. The influence of changes in marital status on developmental trajectories of alcohol use in young adults. *J Stud Alcohol,* 59**,** 647-58.

DAICHES, D. 1978. *Scotch whisky : its past and present / David Daiches; with colour photographs by Alan Daiches,* London, Deutsch.

DALY, C. 2014. *Mental Health Services and Social Inclusion in Remote and Rural Areas of Scotland and Canada: A Qualitative Comparison.* Doctor of Philosophy, University of Aberdeen.

DAVIS, C. N. & O'NEILL, S. E. 2022. Treatment of Alcohol Use Problems Among Rural Populations: a Review of Barriers and Considerations for Increasing Access to Quality Care. *Curr Addict Rep,* 9**,** 432-444.

DEPARTMENT OF HEALTH AND SOCIAL CARE 2011. Physical Activity Guidelines: UK Chief Medical Officers' Report. London, UK.

DEPARTMENT OF HEALTH ENGLAND, W. G., DEPARTMENT OF HEALTH IRELAND, SCOTTISH GOVERNMENT 2016. UK Chief Medical Officers' Low-Risk Drinking Guidelines 2016. London: UK Chief Medical Officers (CMOs),.

DEPARTMENT OF HEATH AND SOCIAL CARE 2016. Alcohol Guidelines Review – Report from the Guidelines Development Group to the UK Chief Medical Officers.

DEVOS-COMBY, L. & LANGE, J. E. 2008. "My drink is larger than yours"? A literature review of self-defined drink sizes and standard drinks. *Current drug abuse reviews,* 1**,** 162-176.

DH/LONG TERM CONDITIONS 2012. Long Term Conditions Compendium of Information. Leeds.

DIAS DA SILVA, D., SILVA, J. P., CARMO, H. & CARVALHO, F. 2021. Neurotoxicity of psychoactive substances: A mechanistic overview. *Current Opinion in Toxicology,* 28**,** 76-83.

DIXON, M. A. & CHARTIER, K. G. 2016. Alcohol Use Patterns Among Urban and Rural Residents: Demographic and Social Influences. *Alcohol Res,* 38**,** 69-77.

DONATH, C., GRASSEL, E., BAIER, D., PFEIFFER, C., KARAGULLE, D., BLEICH, S. & HILLEMACHER, T. 2011. Alcohol consumption and binge drinking in adolescents: comparison of different migration backgrounds and rural vs. urban residence--a representative study. *BMC Public Health,* 11**,** 84.

ELREEDY, D. & ATIYA, A. F. 2019. A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. *Information Sciences,* 505**,** 32-64.

EMSLIE, C., HUNT, K. & LYONS, A. 2015. Transformation and time-out: the role of alcohol in identity construction among Scottish women in early midlife. *Int J Drug Policy,* 26**,** 437-45.

ERSKINE, S., MAHESWARAN, R., PEARSON, T. & GLEESON, D. 2010. Socioeconomic deprivation, urban-rural location and alcohol-related mortality in England and Wales. *BMC Public Health,* 10**,** 99.

ESSER, M. B. & JERNIGAN, D. H. 2018. Policy Approaches for Regulating Alcohol Marketing in a Global Context: A Public Health Perspective. *Annu Rev Public Health,* 39**,** 385-401.

FERNÁNDEZ, A., GARCIA, S., HERRERA, F., CHAWLA, N. 2018. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research,* 61**,** 863-905.

FRIESEN, E. L. & KURDYAK, P. 2020. Alcohol use and alcohol-related harm in rural and remote communities: protocol for a scoping review. *BMJ Open,* 10**,** e036753.

GENUER, R. & POGGI, J. M. 2020. Random Forests. *In:* GENUER, R. & POGGI, J.-M. (eds.) *Random Forests with R.* Cham: Springer International Publishing.

GEOGRAPHIC INFORMATION SCIENCE & ANALYSIS TEAM 2014. Scottish Government urban/rural classification 2013-2014. *In:* RURAL AND ENVIRONMENT SCIENCE AND ANALYTICAL SERVICES DIVISION (ed.). Edinburgh: The Scottish Government,.

GEOGRAPHIC INFORMATION SCIENCE & ANALYSIS TEAM 2018. Scottish Government Urban Rural Classification 2016. *In:* RURAL AND ENVIRONMENT SCIENCE AND ANALYTICAL SERVICES DIVISION (ed.). Edinburgh: The Scottish Government,.

GEOGRAPHIC INFORMATION SCIENCE & ANALYSIS TEAM 2022. Scottish Government Urban Rural Classification 2020. *In:* RURAL AND ENVIRONMENT SCIENCE AND ANALYTICAL SERVICES DIVISION (ed.). Edinburgh: The Scottish Government,.

GILES, L., MACKAY, D., RICHARDSON, E., LEWSEY, J., ROBINSON, M. & BEESTON, C. 2024. Evaluating the impact of minimum unit pricing (MUP) on alcohol sales after 3 years of implementation in Scotland: A controlled interrupted time-series study. *Addiction,* 119**,** 1378-1386.

GILLIGAN, C., ANDERSON, K. G., LADD, B. O., YONG, Y. M. & DAVID, M. 2019. Inaccuracies in survey reporting of alcohol consumption. *BMC Public Health,* 19**,** 1639.

GRANT, B. F., GOLDSTEIN, R. B., SAHA, T. D., CHOU, S. P., JUNG, J., ZHANG, H., PICKERING, R. P., RUAN, W. J., SMITH, S. M., HUANG, B. & HASIN, D. S. 2015. Epidemiology of DSM-5 Alcohol Use Disorder: Results From the National Epidemiologic Survey on Alcohol and Related Conditions III. *JAMA Psychiatry,* 72**,** 757-66.

GRITTNER, U., KUNTSCHE, S., GMEL, G. & BLOOMFIELD, K. 2013. Alcohol consumption and social inequality at the individual and country levels--results from an international study. *Eur J Public Health,* 23**,** 332-9.

HUGHES, J., LIVINGSTON, W., BUYKX, P., JOHNSTON, A., LITTLE, S., MCCARTHY, T., MCLEAN, A., PERKINS, A., WRIGHT, A. & HOLMES, J. 2023. Views on minimum unit pricing for alcohol before its introduction among people with alcohol dependence in Scotland: A qualitative interview study. *Drug Alcohol Rev,* 42**,** 1338-1348.

IBM CORP. 2023. IBM SPSS Statistics for Macintosh. 29.0.2.0 ed. Armonk, New York.

IM, P. K., MILLWOOD, I. Y., GUO, Y., DU, H., CHEN, Y., BIAN, Z., TAN, Y., GUO, Z., WU, S., HUA, Y., LI, L., YANG, L., CHEN, Z. & CHINA KADOORIE BIOBANK COLLABORATIVE, G. 2019. Patterns and trends of alcohol consumption in rural and urban areas of China: findings from the China Kadoorie Biobank. *BMC Public Health,* 19**,** 217.

INSTITUTE OF ALCOHOL STUDIES (IAS) 2020. Alcohol and health inequalities. IAS,.

INTERNATIONAL LABOUR ORGANIZATION (ILO). 1996. *Concepts and definitions* [Online]. ILO Department of Statistics,. Available: https://ilostat.ilo.org/methods/concepts-and-definitions/ [Accessed 01/06/2024].

JAPKOWICZ, N. The class imbalance problem: Significance and strategies. Proc. of the Int'l Conf. on artificial intelligence, 2000. 111-117.

KIRASICH, K. S., T & SADLER, B. Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. 2018.

KLOEP, M., HENDRY, L. B., INGEBRIGTSEN, J. E., GLENDINNING, A. & ESPNES, G. A. 2001. Young people in 'drinking' societies? Norwegian, Scottish and Swedish adolescents' perceptions of alcohol use. *Health Educ Res,* 16**,** 279-91.

KUHN, M. 2008. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software,* 28**,** 1-26.

KWASNICKA, D., BOROUJERDI, M., O'GORMAN, A., ANDERSON, M., CRAIG, P., BOWMAN, L. & MCCANN, M. 2021. An N-of-1 study of daily alcohol consumption following minimum unit pricing implementation in Scotland. *Addiction,* 116**,** 1725-1733.

LAPPAS, N. T. & LAPPAS, C. M. 2022. Ethanol. *In:* LAPPAS, N. T. & LAPPAS, C. M. (eds.) *Forensic Toxicology.* San Diego: Academic Press.

LEDOLTER, J. 2013. *Data Mining and Business Analytics with R,* Somerset, Wiley.

LI, J., LOVATT, M., EADIE, D., DOBBIE, F., MEIER, P., HOLMES, J., HASTINGS, G. & MACKINTOSH, A. M. 2017. Public attitudes towards alcohol control policies in Scotland and England: Results from a mixed-methods study. *Soc Sci Med,* 177**,** 177-189.

LIVINGSTON, W., HOLMES, J., HUGHES, J., BUYKX, P., PERKINS, A., WRIGHT, A., GARDINER, K., YANNOULIS, Y., JOHNSTON, A. & MACLEAN, A. 2023. Expected and actual responses to minimum unit pricing (MUP) for alcohol of people drinking at harmful levels in Scotland. *Drugs-Education Prevention and Policy***,** 1-10.

LÓPEZ, V., FERNÁNDEZ, A.,GARCÍA, S., PALADE, V., HERRERA, F. 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information sciences,* 250**,** 113-141.

LOUPPE, G. 2014. *Understanding Random Forests: From Theory to Practice.*

LOUPPE, G. 2015. *Understanding Random Forests: From Theory to Practice.* PhD, Université de Liège.

MALEK-AHMADI, M. & DEGIORGIO, L. 2015. Risk of alcohol abuse in urban versus rural DUI offenders. *Am J Drug Alcohol Abuse,* 41**,** 353-7.

MANCA, F., PARAB, R., MACKAY, D., FITZGERALD, N. & LEWSEY, J. 2024a. Evaluating the impact of minimum unit pricing for alcohol on road traffic accidents in Scotland after 20 months: An interrupted time series study. *Addiction,* 119**,** 509-517.

MANCA, F., ZHANG, L., FITZGERALD, N., HO, F., INNES, H., JANI, B., KATIKIREDDI, S. V., MCAULEY, A., SHARP, C. & LEWSEY, J. 2024b. Pharmacological treatments for alcohol dependence: Evidence on uptake, inequalities and comparative effectiveness from a UK population-based cohort. *Drug Alcohol Rev,* 43**,** 1183-1193.

MARTIN, G., INCHLEY, J. & CURRIE, C. 2019a. Do Drinking Motives Mediate the Relationship between Neighborhood Characteristics and Alcohol Use among Adolescents? *Int J Environ Res Public Health,* 16**,** 853.

MARTIN, G., INCHLEY, J., MARSHALL, A., SHORTT, N. & CURRIE, C. 2019b. The neighbourhood social environment and alcohol use among urban and rural Scottish adolescents. *Int J Public Health,* 64**,** 95-105.

MARTIN, G. D. M., I. 2008. *The invention of whiskey / by Graham Dunstan Martin ; illustrations by Iain McIntosh after George Bain after the Book of Kells,* Edinburgh, Maclean Dubois.

MCCAUL, M. E., ROACH, D., HASIN, D. S., WEISNER, C., CHANG, G. & SINHA, R. 2019. Alcohol and Women: A Brief Overview. *Alcohol Clin Exp Res,* 43**,** 774-779.

MCDONALD, M. 1994. *Gender, drink and drugs,* Oxford, Berg Publishers.

MCGOVERN, P. E. 2009. *Uncorking the Past: The Quest for Wine, Beer, and Other Alcoholic Beverages.*, University of California Press.

MCHUGH, M. L. 2013. The Chi-square test of independence. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara /,* 23**,** 143-149.

MICHALAK, L., TROCKI, K. & BOND, J. 2007. Religion and alcohol in the U.S. National Alcohol Survey: how important is religion for abstention and drinking? *Drug Alcohol Depend,* 87**,** 268-80.

MILLER, P. G., COOMBER, K., STAIGER, P., ZINKIEWICZ, L. & TOUMBOUROU, J. W. 2010. Review of rural and regional alcohol research in Australia. *Aust J Rural Health,* 18**,** 110-7.

MOHAMED, S. & AJMAL, M. 2015. Multivariate analysis of binge drinking in young adult population: Data analysis of the 2007 Survey of Lifestyle, Attitude and Nutrition in Ireland. *Psychiatry Clin Neurosci,* 69**,** 483-8.

MORRIS, H., LARSEN, J., CATTERALL, E., MOSS, A. C. & DOMBROWSKI, S. U. 2020. Peer pressure and alcohol consumption in adults living in the UK: a systematic qualitative review. *BMC Public Health,* 20**,** 1014.

MUCHLINSKI, D., SIROKY, D., HE, J. R. & KOCHER, M. 2016. Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. *Political Analysis,* 24**,** 87-103.

NATIONAL HEALTH SERVICE (NHS). 2021. *Age: How to talk about different age groups and stages of life.* [Online]. National Health Service (NHS),. Available: https://service-manual.nhs.uk/content/inclusive-content/age [Accessed 01/06/2024].

NATIONAL INSTITUTE ON ALCOHOL ABUSE AND ALCOHOLISM (NIAAA). *Alcohol's Effects on Health: Alcohol Topics A to Z* [Online]. NIDA,. Available: https://www.niaaa.nih.gov/alcohols-effects-health/alcohol-topics-a-to-z [Accessed 15/06/2024 2024].

NATIONAL INSTITUTE ON ALCOHOL ABUSE AND ALCOHOLISM (NIAAA) 2024. National Institute on Alcohol Abuse and Alcoholism Strategic Plan: Fiscal Years 2024–2028. Advancing Alcohol Research to Promote Health and Well-Being.: NIAAA,.

NATIONAL RECORDS OF SCOTLAND (NRS) 2023. Alcohol-specific Deaths in Scotland 2022. NRS,.

NIKULIN, M. S. & CHIMITOVA, E. V. 2017. *Chi-squared Goodness-of-fit Tests for Censored Data,* London, UK, ISTE, Ltd.

NUGAWELA, M. D., LANGLEY, T., SZATKOWSKI, L. & LEWIS, S. 2016. Measuring Alcohol Consumption in Population Surveys: A Review of International Guidelines and Comparison with Surveys in England. *Alcohol Alcohol,* 51**,** 84-92.

OFFICE OF NATIONAL STATISTICS. n.d. *The National Statistics Socio-economic classification (NS-SEC)* [Online]. Office of National Statistics,. Available: https://www.ons.gov.uk/methodology/classificationsandstandards/otherclassifications/thenationalstatisticssocioeconomicclassificationnssecrebasedonsoc2010 [Accessed 01/06/2024].

OHANA-LEVI, N. & NETZER, Y. 2023. Long-Term Trends of Global Wine Market. *Agriculture-Basel,* 13**,** 224.

PENG, C. Y. J., LEE, K. L. & INGERSOLL, G. M. 2002. An introduction to logistic regression analysis and reporting. *Journal of Educational Research,* 96**,** 3-14.

PHILLIPS, R. 2014. *Alcohol*, University of North Carolina Press.
PROBST, P., WRIGHT, M. N. & BOULESTEIX, A. L. 2019. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery,* 9**,** e1301.
PUBLIC HEALTH SCOTLAND (PHS) 2024. Alcohol Related Hospital Statistics Scotland 2022/23. PHS,.
R DEVELOPMENT CORE TEAM 2024. R: A language and environment for statistical computing. Vienna, Austria: Foundation for Statistical Computing.
RANSTAM, J. & COOK, J. A. 2018. LASSO regression. *British Journal of Surgery,* 105**,** 1348-1348.
RICHARDSON, E. A., HILL, S. E., MITCHELL, R., PEARCE, J. & SHORTT, N. K. 2015. Is local alcohol outlet density related to alcohol-related morbidity and mortality in Scottish cities? *Health Place,* 33**,** 172-80.
RIGATTI, S. J. 2017. Random Forest. *J Insur Med,* 47**,** 31-39.
RITCHIE, C. 2007. Beyond drinking: the role of wine in the life of the UK consumer. *International Journal of Consumer Studies,* 31**,** 534-540.
ROMBERGER, D. J. & GRANT, K. 2004. Alcohol consumption and smoking status: the role of smoking cessation. *Biomed Pharmacother,* 58**,** 77-83.
SCOTCEN SOCIAL RESEARCH 2016. Consultation on the questionnaire content of the Scottish Health Survey. Edinburgh.
SCOTCEN SOCIAL RESEARCH 2017a. Questionnaire Content of the Scottish Health Survey. Consultation Analysis Report. Edinburgh.
SCOTCEN SOCIAL RESEARCH 2017b. The Scottish Health Survey: Main Report. *In:* THE SCOTTISH GOVERNMENT (ed.) *The Scottish Health Survey.* 2017 ed. Edinburgh: The Scottish Government,.
SCOTCEN SOCIAL RESEARCH 2017c. The Scottish Health Survey: Technical Report. *In:* THE SCOTTISH GOVERNMENT (ed.) *The Scottish Health Survey.* 2017 ed. Edinburgh: The Scottish Government,.
SCOTCEN SOCIAL RESEARCH 2018a. The Scottish Health Survey: Main Report. *In:* THE SCOTTISH GOVERNMENT (ed.) *The Scottish Health Survey.* 2018 ed. Edinburgh: The Scottish Government,.
SCOTCEN SOCIAL RESEARCH 2018b. The Scottish Health Survey: Technical Report. *In:* THE SCOTTISH GOVERNMENT (ed.) *The Scottish Health Survey.* 2018 ed. Edinburgh: The Scottish Government,.
SCOTCEN SOCIAL RESEARCH 2019a. The Scottish Health Survey: Main Report. *In:* THE SCOTTISH GOVERNMENT (ed.) *The Scottish Health Survey.* 2019 ed. Edinburgh: The Scottish Government,.
SCOTCEN SOCIAL RESEARCH 2019b. The Scottish Health Survey: Technical Report. *In:* THE SCOTTISH GOVERNMENT (ed.) *The Scottish Health Survey.* 2019 ed. Edinburgh: The Scottish Government,.
SCOTCEN SOCIAL RESEARCH 2021a. Scottish Health Survey combined dataset 2017 to 2021: User Guide. *In:* UK DATA ARCHIVE (ed.) *Scottish Health Survey.* London: ScotCen Social Research,.
SCOTCEN SOCIAL RESEARCH 2021b. The Scottish Health Survey: Main Report. *In:* THE SCOTTISH GOVERNMENT (ed.) *The Scottish Health Survey.* 2021 ed. Edinburgh: The Scottish Government,.
SCOTCEN SOCIAL RESEARCH 2021c. The Scottish Health Survey: Technical Report. *In:* THE SCOTTISH GOVERNMENT (ed.) *The Scottish Health Survey.* 2021 ed. Edinburgh: The Scottish Government.
SCOTCEN SOCIAL RESEARCH 2022a. The Scottish Health Survey: Main Report. *In:* THE SCOTTISH GOVERNMENT (ed.) *The Scottish Health Survey.* 2022 ed. Edinburgh: The Scottish Government,.
SCOTCEN SOCIAL RESEARCH 2022b. The Scottish Health Survey: Technical Report. *In:* THE SCOTTISH GOVERNMENT (ed.) *The Scottish Health Survey.* 2022 ed. Edinburgh: The Scottish Government,.
SCOTCEN SOCIAL RESEARCH 2023. Scottish Health Survey, 2021 *In:* UK DATA SERVICE (ed.) 2nd Edition ed.
SCOTTISH HEALTH ACTION ON ALCOHOL PROBLEMS (SHAAP) 2020. Rural Matters. Understanding alcohol use in rural Scotland: Findings from a qualitative research study. SHAAP,.
SCOTTISH HEALTH ACTION ON ALCOHOL PROBLEMS (SHAAP) 2021. Manifesto for the 2021 Scottish Parliament Election.: SHAAP,.
SHIUE, I. 2014. Modeling the effects of indoor passive smoking at home, work, or other households on adult cardiovascular and mental health: the Scottish Health Survey, 2008-2011. *Int J Environ Res Public Health,* 11**,** 3096-107.
SHORTT, N. K., RIND, E., PEARCE, J., MITCHELL, R. & CURTIS, S. 2018. Alcohol risk environments, vulnerability and social inequalities in alcohol consumption. *Ann Am Assoc Geogr,* 108**,** 1210-1227.
SHORTT, N. K., TISCH, C., PEARCE, J., MITCHELL, R., RICHARDSON, E. A., HILL, S. & COLLIN, J. 2015. A cross-sectional analysis of the relationship between tobacco and alcohol outlet density and neighbourhood deprivation. *BMC Public Health,* 15**,** 1014.
SMITH, L. & FOXCROFT, D. 2009. Drinking in the UK. *An exploration of trends. York: Joseph Rowntree Foundation*.

SPOERRI, A., ZWAHLEN, M., PANCZAK, R., EGGER, M., HUSS, A. & SWISS NATIONAL, C. 2013. Alcohol-selling outlets and mortality in Switzerland--the Swiss National Cohort. *Addiction,* 108**,** 1603-11.
STILL, H. 2024. Alcohol and Other Drug Harm in Rural Scotland: Addiction Looks Different Here. *In:* TURBETT, C. & PYE, J. (eds.) *Rural Social Work in the UK.* Cham: Springer International Publishing.
STOCKWELL, T., ZHAO, J., MACDONALD, S., VALLANCE, K., GRUENEWALD, P., PONICKI, W., HOLDER, H. & TRENO, A. 2011. Impact on alcohol-related mortality of a rapid rise in the density of private liquor outlets in British Columbia: a local area multi-level analysis. *Addiction,* 106**,** 768-76.
STOLTZFUS, J. C. 2011. Logistic regression: a brief primer. *Acad Emerg Med,* 18**,** 1099-104.
TECKLE, P., HANNAFORD, P. & SUTTON, M. 2012. Is the health of people living in rural areas different from those in cities? Evidence from routine data linked with the Scottish Health Survey. *BMC Health Serv Res,* 12**,** 43.
THE SCOTTISH GOVERNMENT 2023. Scottish Crime and Justice Survey 2021/22: Main Findings. *In:* THE SCOTTISH GOVERNMENT (ed.). Edinburgh: The Scottish Government,.
THE SCOTTISH GOVERNMENT. n.d.-a. *Scottish Health Survey.* [Online]. The Scottish Government,. Available: https://www.gov.scot/collections/scottish-health-survey/ [Accessed 12/04/2024].
THE SCOTTISH GOVERNMENT. n.d.-b. *Scottish Index of Multiple Deprivation 2020.* [Online]. Edinburgh. Available: https://www.gov.scot/collections/scottish-index-of-multiple-deprivation-2020/ [Accessed 01/06/2024].
TORNEY, A., ROOM, R., JIANG, H., HUCKLE, T., HOLMES, J. & CALLINAN, S. 2024. Where do high-risk drinking occasions occur more often? A cross-sectional, cross-country study. *Drug Alcohol Rev,* 43**,** 1172-1177.
UK DATA SERVICE. 2024. *Scottish Health Survey Serie* [Online]. Available: https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=2000047#!/access-data [Accessed 12/04/2024].
VAN OERS, J. A., BONGERS, I. M., VAN DE GOOR, L. A. & GARRETSEN, H. F. 1999. Alcohol consumption, alcohol-related problems, problem drinking, and socioeconomic status. *Alcohol Alcohol,* 34**,** 78-88.
VENABLES, W. N. & SMITH, D. M. 2024. An Introduction to R. Version 4.4.1.
WICKHAM, H. 2019. *Advanced R,* Boca Raton, CRC Press/Taylor & Francis Group.
WICKHAM, H., ÇETINKAYA-RUNDEL, M. & GROLEMUND, G. 2023. *R for data science : import, tidy, transform, visualize, and model data,* Beijing, O'Reilly.
WILKINSON, L. 2005. *The grammar of graphics,* New York, Springer.
WORLD HEALTH ORGANISATION (WHO) 2010. Global strategy to reduce the harmful use of alcohol. Geneva: WHO,.
WORLD HEALTH ORGANISATION (WHO) 2018. Global status report on alcohol and health. Geneva: WHO,.
WORLD HEALTH ORGANISATION (WHO) 2023. Global alcohol action plan 2022-2030 (Pre-printed copy). Geneva: WHO,.
WRIGHT, M. N. & ZIEGLER, A. 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C plus plus and R. *Journal of Statistical Software,* 77**,** 1-17.
YAMASHITA, T., YAMASHITA, K. & KAMIMURA, R. 2007. A stepwise AIC method for variable selection in linear regression. *Communications in Statistics-Theory and Methods,* 36**,** 2395-2403.
ZHOU, T., SUN, D., LI, X., MA, H., HEIANZA, Y. & QI, L. 2021. Educational attainment and drinking behaviors: Mendelian randomization study in UK Biobank. *Mol Psychiatry,* 26**,** 4355-4366.

APPENDIX: SUPPLEMENTARY TABLES

LIST OF SUPPLEMENTARY TABLES

TABLE I.1 PROPORTION OF CATEGORY'S VARIABLES IN URBAN & RURAL, URBAN AND RURAL DATASETS 82

TABLE I.2 ALCOHOL CONSUMPTION BELOW OR ABOVE THE WEEKLY LIMIT IN URBAN & RURAL, URBAN AND RURAL DATASETS 84

TABLE I.3 COEFICIENTS AFTER CROSS-VALIDATION IN THE URBAN & RURAL TRAINING DATASET (DW.TRAIN) 88

TABLE I.4 COEFICIENTS AFTER CROSS-VALIDATION IN THE URBAN & RURAL OVERSAMPLED TRAINING DATASET - RATIO 30% (UP30DW.TRAIN) 89

TABLE I.1 COEFICIENTS AFTER CROSS-VALIDATION IN THE URBAN & RURAL OVERSAMPLED TRAINING DATASET - RATIO 40% (UP40DW.TRAIN) 90

TABLE I.2 COEFICIENTS AFTER CROSS-VALIDATION IN THE URBAN & RURAL OVERSAMPLED TRAINING DATASET – RATIO 50% (UP50DW.TRAIN)..... 91

TABLE I.7 COEFICIENTS AFTER CROSS-VALIDATION IN THE SMOTE URBAN & RURAL TRAINING DATASET K = 3 / OR = 0.5 (SMK3R50_DW) 92

TABLE I.8 COEFICIENTS AFTER CROSS-VALIDATION IN THE SMOTE URBAN & RURAL TRAINING DATASET K = 3 / OR = 0.75 (SMK3R75_DW) 94

TABLE I.9 COEFICIENTS AFTER CROSS-VALIDATION IN THE SMOTE URBAN & RURAL TRAINING DATASET K = 3 / OR = 1 (SMK3R1_DW) 95

TABLE I.10 COEFICIENTS AFTER CROSS-VALIDATION IN THE SMOTE URBAN & RURAL TRAINING DATASET K = 5 / OR = 0.5 (SMK5R50_DW) 96

TABLE I.11 COEFICIENTS AFTER CROSS-VALIDATION IN THE SMOTE URBAN & RURAL TRAINING DATASET K = 5 / OR = 0. 75 (SMK5R75_DW)..... 98

TABLE I.12 COEFICIENTS AFTER CROSS-VALIDATION IN THE SMOTE URBAN & RURAL TRAINING DATASET K = 5 / OR = 1 (SMK5R100_DW) 100

TABLE I.13 COEFICIENTS AFTER CROSS-VALIDATION IN THE SMOTE URBAN & RURAL TRAINING DATASET K = 10 / OR = 0. 5 (SMK10R50_DW)..... 102

TABLE I.14 COEFICIENTS AFTER CROSS-VALIDATION IN THE SMOTE URBAN & RURAL TRAINING DATASET K = 10 / OR = 0. 75 (SMK10R75_DW)..... 104

TABLE I.15 COEFICIENTS AFTER CROSS-VALIDATION IN THE SMOTE URBAN & RURAL TRAINING DATASET K = 10 / OR = 1 (SMK10R100_DW) 105

TABLE I.16 COEFICIENTS AFTER CROSS-VALIDATION IN THE TRAINING URBAN DATASET (DWU.TRAIN) 107

TABLE I.17 COEFICIENTS FROM CROSS-VALIDATION IN THE OVERSAMPLING URBAN TRAINING DATASET 30% (UP30_DWU) 108

TABLE I.18 COEFICIENTS FROM CROSS-VALIDATION IN THE OVERSAMPLING URBAN TRAINING DATASET 40% (UP40_DWU) 109

TABLE I.19 COEFICIENTS FROM CROSS-VALIDATION IN THE OVERSAMPLING URBAN TRAINING DATASET 50% (UP50_DWU) 110

TABLE I.20 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE URBAN TRAINING DATASET K = 3 OR = 0.5 (SMK3R50_DWU)..... 111

TABLE I.21 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE URBAN TRAINING DATASET K = 3 OR = 0.75 (SMK3R75_DWU)..... 112

TABLE I.22 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE URBAN TRAINING DATASET K = 3 OR = 1 (SMK3R1_DWU)..... 113

TABLE I.23 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE URBAN TRAINING DATASET K = 5 OR = 0.5 (SMK5R50_DWU)..... 114

TABLE I.24 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE URBAN TRAINING DATASET K = 5 OR = 0.75 (SMK5R0.75_DWU)..... 115

TABLE I.25 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE URBAN TRAINING DATASET K = 5 OR = 1 (SMK5R100_DWU)..... 116

TABLE I.26 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE URBAN TRAINING DATASET K = 10 OR = 0.5 (SMK10R50_DWU)..... 117

TABLE I.27 COEFICIENTS FROM CROSS-VALIDATION IN THE SMOTE URBAN TRAINING DATASET K = 10 OR = 0.75 (SMK10R75_DWU)..... 118

TABLE I.28 COEFICIENTS FROM CROSS-VALIDATION IN THE SMOTE URBAN TRAINING DATASET K = 10 OR = 1 (SMK10R100_DWU)..... 119

TABLE I.29 COEFICIENTS FROM CROSS-VALIDATION IN THE RURAL TRAINING DATASET (DWR_TRAIN)..... 121

TABLE I.30 COEFICIENTS FROM CROSS-VALIDATION IN THE OVERSAMPLING RURAL TRAINING DATASET 30% (UP30_DWR)..... 122

TABLE I.31 COEFICIENTS FROM CROSS-VALIDATION IN THE OVERSAMPLING RURAL TRAINING DATASET 40% (UP40_DWR)..... 123

TABLE I.32 COEFICIENTS FROM CROSS-VALIDATION IN THE OVERSAMPLING RURAL TRAINING DATASET 50% (UP50_DWR)..... 124

TABLE I.33 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE TRAINING RURAL DATASET K = 3 / OR = 0.5 (SMK3R50_DWR)..... 126

TABLE I.34 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE TRAINING RURAL DATASET K = 3 / OR = 0.75 (SMK3R75_DWR)..... 127

TABLE I.35 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE TRAINING RURAL DATASET K = 3 / OR = 1 (SMK3R100_DWR)..... 128

TABLE I.36 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE TRAINING RURAL DATASET K = 5 / OR = 0.5 (SMK5R50_DWR)..... 129

TABLE I.37 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE TRAINING RURAL DATASET K = 5 / OR = 0.75 (SMK5R75_DWR)..... 130

TABLE I.38 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE TRAINING RURAL DATASET K = 5 / OR = 1 (SMK5R100_DWR)..... 131

TABLE I.39 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE TRAINING RURAL DATASET K = 10 / OR = 0.50 (SMK10R50_DWR)..... 133

TABLE I.40 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE TRAINING RURAL DATASET K = 10 / OR = 0.75 (SMK10R75_DWR)..... 134

TABLE I.41 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE TRAINING RURAL DATASET K = 10 / OR = 1 (SMK10R100_DWR)..... 135

TABLE I.42 ESTIMATES AND OR WITH 95% CI OF THE FINAL MODEL FOR THE URBAN & RURAL DATASET..... 136

TABLE I.43 ESTIMATES AND OR WITH 95% CI OF THE FINAL MODEL FOR THE URBAN DATASET..... 138

TABLE I.44 ESTIMATES AND OR WITH 95% CI OF THE FINAL MODEL FOR THE RURAL DATASET..... 140

TABLE I.1 PROPORTION OF CATEGORY'S VARIABLES IN URBAN & RURAL, URBAN AND RURAL DATASETS

Variable	Population Classification		
	Urban & Rural	Urban	Rural
Survey Year (syear)	N = 11,185	N = 8,965	N = 2,220
2017	2,353 (21.54%)	1,864 (20.79%)	489 (22.03%)
2018	2,995 (26.78%)	2,412 (26.90%)	583 (26.26%)
2019	3,046 (27.23%)	2,453 (27.36%)	593 (26.71%)
2021	2,791 (24.95%)	2,236 (24.94%)	555 (25.00%)
Sex of respondent (sex)	N = 11,185	N = 8,965	N = 2,220
Male	4,722 (42.22%)	3,794 (42.32%)	928 (41.80%)
Female	6,463 (57.78%)	5,171 (57.68%)	1,292 (58.20%)
Age (ag16g10)	N = 11,185	N = 8,965	N = 2,220
25-34	2,183 (19.52%)	1,873 (20.89%)	310 (13.96%)
35-44	2,515 (22.49%)	2,030 (22.64%)	485 (21.85%)
45-54	3,037 (27.15%)	2,390 (26.66%)	647 (29.14%)
55-64	3,450 (30.84%)	2,672 (29.80%)	778 (35.05%)
Birthplace (birthpla3)	N = 11,156	N = 8,939	N = 2,217
Scotland	8,388 (75.19%)	6,844 (76.56%)	1,544 (69.64%)
Rest of the UK	1,575 (14.12%)	1,051 (11.76%)	524 (23.64%)
Elsewhere	1,193 (10.69%)	1,044 (11.68%)	149 (6.72%)
Ethnicity (ethnic05)	N = 11,150	N = 8,933	N = 2,217
White: Scottish	8,394 (75.28%)	6,804 (76.17%)	1,590 (71.72%)
White: Rest of the UK	1,568 (14.06%)	1,070 (11.98%)	498 (22.46%)
White: Other	690 (6.19%)	586 (6.56%)	104 (4.69%)
Asian	304 (2.73%)	291 (3.26%)	13 (0.59%)
Other minority ethnic	194 (1.74%)	182 (2.04%)	12 (0.54%)
Religion (religi04)	N = 11,145	N = 8,931	N = 2,214
None	6,127 (54.98%)	4,853 (54.34%)	1,274 (57.54%)
Church of Scotland	2,262 (20.30%)	1,751 (19.61%)	511 (23.08%)
Roman Catholic	1,508 (13.53%)	1,329 (14.88%)	179 (8.08%)
Other Christian	893 (8.01%)	667 (7.47%)	226 (10.21%)
Other religions	355 (3.19%)	331 (3.71%)	24 (1.08%)
Marital status (maritalg)	N = 11,183	N = 8,963	N = 2,220
Married / civil partnership	6,045 (54.06%)	4,656 (51.95%)	1,389 (62.57%)
Living as married	1,705 (15.25%)	1,375 (15.34%)	330 (14.86%)
Single	2,046 (18.30%)	1,768 (19.73%)	278 (12.52%)
Separated	346 (3.09%)	292 (3.26%)	54 (2.43%)
Divorced or dissolved	842 (7.53%)	714 (7.97%)	128 (5.77%)
Widowed	199 (1.78%)	158 (1.76%)	41 (1.85%)
Life satisfaction (lifesat2)	N = 11,160	N = 8,943	N = 2,217
Above mode (mode = 8)	3,763 (33.72%)	2,891 (32.33%)	872 (39.33%)
Mode (mode = 8)	3,492 (31.29%)	2,778 (31.06%)	714 (32.21%)
Below mode (mode = 8)	3,905 (34.99%)	3,274 (36.61%)	631 (28.46%)
General health (genhelf)	N = 11,180	N = 8,961	N = 2,219
Very good	3,928 (35.13%)	3,089 (34.47%)	839 (37.81%)
Good	4,397 (39.33%)	3,506 (39.13%)	891 (40.15%)
Fair	1,930 (17.26%)	1,582 (17.65%)	348 (15.68%)
Bad	693 (6.20%)	586 (6.54%)	107 (4.82%)
Very bad	232 (2.08%)	198 (2.21%)	34 (1.53%)
LTC (limitac_h)	N = 11,181	N = 8,962	N = 2,219
No limitations	6,205 (55.50%)	4,931 (55.02%)	1,274 (57.41%)
Not at all	1,486 (13.29%)	1,175 (13.11%)	311 (14.02%)
A little	1,676 (14.99%)	1,345 (15.01%)	331 (14.92%)
A lot	1,814 (16.22%)	1,511 (16.86%)	303(13.65%)
Physical activity level (adt10gpW)	N = 11,121	N = 8,912	N = 2,209
Meets recommendations	7,870 (70.77%)	6,231 (69.92%)	1,639 (74.20%)
Some activity	1,103 (9.92%)	903 (10.13%)	200 (9.05%)
Low activity	394 (3.54%)	318 (3.57%)	76 (3.44%)
Very low activity	1,754 (15.77%)	1,460 (16.38%)	294 (13.31%)
Smoking (cig)	N = 11,097	N = 8,895	N = 2,202
Never smoked	5,795 (52.22%)	4,642 (52.19%)	1,153 (52.36%)
Ex-smoker	3,362 (30.30%)	2,631 (29.58%)	731 (33.20%)
Light smokers	677 (6.10%)	578 (6.50%)	99 (4.50%)
Moderate smokers	821 (7.40%)	676 (7.60%)	145 (6.58%)
Heavy smokers	442 (3.98%)	368 (4.14%)	74 (3.36%)
Educational level (hedqul08)	N = 11,137	N = 8,923	N = 2,214
Degree or higher	4,812 (43.21%)	3,814 (42.74%)	998 (45.08%)
HNC/D	1,646 (14.78%)	1,327 (14.87%)	319 (14.41%)
Higher grade or equivalent	1,580 (14.19%)	1,232 (13.81%)	348 (15.72%)
Standard grade or equivalent	1,890 (16.97%)	1,553 (17.40%)	337 (15.22%)
Other school level	188 (1.69%)	167 (1.87%)	21 (0.95%)
No qualifications	1,021 (9.17%)	830 (9.30%)	191 (8.63%)

TABLE I.1 PROPORTION OF CATEGORY'S VARIABLES IN URBAN & RURAL, URBAN AND RURAL DATASETS

Variable	Population Classification		
	Urban & Rural	Urban	Rural
Social Class (schrpg7)	N = 11,029	N = 8,833	N = 2,196
I Professional	1,091 (9.89%)	880 (9.96%)	211 (9.61%)
II Managerial technical	4,380 (39.71%)	3,429 (38.82%)	951 (43.31%)
IIIN Skilled non-manual	1,517 (13.75%)	1,280 (14.49%)	237 (10.79%)
IIIM Skilled manual	1,896 (17.19%)	1,506 (17.05%)	390 (17.76%)
IV Semi-skilled manual	1,463 (13.27%)	1,173 (13.28%)	290 (13.21%)
V Unskilled manual	467 (4.23%)	375 (4.25%)	92 (4.19%)
Others	215 (1.95%)	190 (2.15%)	25 (1.14%)
Economic activity (neconacb)	N = 11,165	N = 8,947	N = 2,218
In employment	8,446 (75.65%)	6,698 (74.86%)	1,748 (78.81%)
ILO unemployed	267 (2.39%)	228 (2.55%)	39 (1.76%)
Inactive	2,452 (21.96%)	2,021 (22.59%)	431 (19.43%)
SIMD quintile (simd20_rpa)	N = 11,185	N = 8,965	N = 2,220
Least deprived	2,324 (20.78%)	2,110 (23.54%)	214 (9.64%)
4th	2,457 (21.97%)	1,576 (17.58%)	881 (39.68%)
3rd	2,315 (20.70%)	1,455 (16.23%)	860 (38.74%)
2nd	2,203 (19.70%)	1,989 (22.19%)	214 (9.64%)
Most deprived	1,886 (16.86%)	1,835 (20.47%)	51 (2.30%)
Total Units of alcohol/week (drating)			
Mean (SD)	10.15	10.21 (18.34)	9.90 (16.26)
Median (IQR)	4.50 (0.35, 13.57)	4.35 (0.33, 13.59)	4.67 (0.49, 13.50)
Range	0.00 - 412.50	0.00 - 412.50	0.00 - 302.50
Units of normal beer/week (nberwu)			
Mean (SD)	3.48	3.59 (10.79)	3.00 (7.82)
Median (IQR)	0.06 (0.00, 2.25)	0.04 (0.00, 2.25)	0.06 (0.00, 2.18)
Range	0.00 – 280.00	0.00 - 280.00	0.00 - 105.00
Units of strong beer/week (sberwu)			
Mean (SD)	0.26	0.30 (4.71)	0.09 (0.94)
Median (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)
Range	0.00 – 210.00	0.00 - 210.00	0.00 - 24.00
Units of spirits/week (spirwu)			
Mean (SD)	2.16	2.21 (7.90)	1.96 (5.66)
Median (IQR)	0.23 (0.00, 1.50)	0.23 (0.00, 1.61)	0.23 (0.00, 1.50)
Range	0.00 – 280.00	0.00 - 280.00	0.00 - 154.00
Units of sherry/week (sherwu)			
Mean (SD)	0.09	0.09 (1.30)	0.08 (0.83)
Median (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)
Range	0.00 – 56.00	0.00 - 56.00	0.00 - 30.00
Units of wine/week (winewu)			
Mean (SD)	4.14	3.99 (8.40)	4.73 (10.85)
Median (IQR)	0.35 (0.00, 5.25)	0.26 (0.00, 4.50)	0.69 (0.00, 6.00)
Range	0.00 – 297.00	0.00 - 126.00	0.00 - 297.00
Units of alcopops/week (popswu)			
Mean (SD)	0.06	0.06 (0.81)	0.03 (0.53)
Median (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)
Range	0.00 - 36.00	0.00 - 36.00	0.00 - 22.50

TABLE I.2 ALCOHOL CONSUMPTION BELOW OR ABOVE THE WEEKLY LIMIT IN URBAN & RURAL, URBAN AND RURAL DATASETS

Variable	Urban & rural dataset			Urban dataset			Rural dataset		
	Above the limit		p-value[2]	Above the limit		p-value[2]	Above the limit		p-value[2]
	No	Yes		No	Yes		No	Yes	
	N = 8,444	N = 2,669		N = 6,767	N = 2,139		N = 1,677	N = 530	
Survey year (syear)			0.71			0.62			0.26
2017	1,776 (75.99%)	561 (24.01%)		1,392 (75.16%)	460 (24.84%)		384 (79.18%)	101 (20.82%)	
2018	2,269 (76.29%)	705 (23.71%)		1,831 (76.48%)	563 (23.52%)		438 (75.52%)	142 (24.48%)	
2019	2,273 (75.24%)	748 (24.76%)		1,837 (75.53%)	595 (24.47%)		436 (74.02%)	153 (25.98%)	
2021	2,126 (76.45%)	655 (23.55%)		1,707 (76.62%)	521 (23.38%)		419 (75.77%)	134 (24.23%)	
Age (ag16g10)			<0.001			<0.001			0.010
25-34	1,772 (81.81%)	394 (18.19%)		1,520 (81.72%)	340 (18.28%)		252 (82.35%)	54 (17.65%)	
35-44	1,972 (79.16%)	519 (20.84%)		1,597 (79.45%)	413 (20.55%)		375 (77.96%)	106 (22.04%)	
45-54	2,229 (73.86%)	789 (26.14%)		1,758 (74.02%)	617 (25.98%)		471 (73.25%)	172 (26.75%)	
55-64	2,471 (71.87%)	967 (28.13%)		1,892 (71.10%)	769 (28.90%)		579 (74.52%)	198 (25.48%)	
Sex of respondent (sex)			<0.001			<0.001			<0.001
Male	3,181 (67.88%)	1,505 (32.12%)		2,549 (67.70%)	1,216 (32.30%)		632 (68.62%)	289 (31.38%)	
Female	5,263 (81.89%)	1,164 (18.11%)		4,218 (82.05%)	923 (17.95%)		1,045 (81.26%)	241 (18.74%)	
Country of birth (birthpla3)			<0.001			<0.001			0.008
Scotland	6,246 (74.81%)	2,103 (25.19%)		5,089 (74.70%)	1,724 (25.30%)		1,157 (75.33%)	379 (24.67%)	
Rest of the UK	1,167 (74.28%)	404 (25.72%)		775 (73.95%)	273 (26.05%)		392 (74.95%)	131 (25.05%)	
Elsewhere	1,027 (86.45%)	161 (13.55%)		899 (86.44%)	141 (13.56%)		128 (86.49%)	20 (13.51%)	
Ethnicity (rthnic05)			<0.001			<0.001			0.007
White: Scottish	6,236 (74.64%)	2,119 (25.36%)		5,034 (74.32%)	1,739 (25.68%)		1,202 (75.98%)	380 (24.02%)	
White: rest the UK	1,158 (74.04%)	406 (25.96%)		796 (74.60%)	271 (25.40%)		362 (72.84%)	135 (27.16%)	
White: Other	580 (84.55%)	106 (15.45%)		490 (84.05%)	93 (15.95%)		90 (87.38%)	13 (12.62%)	
Asian	291 (95.72%)	13 (4.28%)		279 (95.88%)	12 (4.12%)		12 (92.31%)	1 (7.69%)	
Other minority ethnic	170 (88.08%)	23 (11.92%)		159 (87.85%)	22 (12.15%)		11 (91.67%)	1 (8.33%)	
Religion (religi04)			<0.001			<0.001			0.004
None	4,549 (74.56%)	1,552 (25.44%)		3,605 (74.64%)	1,225 (25.36%)		944 (74.27%)	327 (25.73%)	
Church of Scotland	1,657 (73.48%)	598 (26.52%)		1,265 (72.45%)	481 (27.55%)		392 (77.01%)	117 (22.99%)	
Roman Catholic	1,168 (77.97%)	330 (22.03%)		1,039 (78.59%)	283 (21.41%)		129 (73.30%)	47 (26.70%)	
Other Christian	725 (81.64%)	163 (18.36%)		538 (81.02%)	126 (18.98%)		187 (83.48%)	37 (16.52%)	
Another religion	332 (93.52%)	23 (6.48%)		309 (93.35%)	22 (6.65%)		23 (95.83%)	1 (4.17%)	
Marital status (maritalg)			0.002			0.008			0.31
Married / civil partnership	4,513 (75.07%)	1,499 (24.93%)		3,482 (75.17%)	1,150 (24.83%)		1,031 (74.71%)	349 (25.29%)	
Living as married	1,260 (74.25%)	437 (25.75%)		1,011 (73.90%)	357 (26.10%)		249 (75.68%)	80 (24.32%)	
Single	1,600 (79.01%)	425 (20.99%)		1,384 (79.13%)	365 (20.87%)		216 (78.26%)	60 (21.74%)	
Separated	263 (76.45%)	81 (23.55%)		219 (75.26%)	72 (24.74%)		44 (83.02%)	9 (16.98%)	
Divorced or dissolved	648 (77.70%)	186 (22.30%)		545 (77.20%)	161 (22.80%)		103 (80.47%)	25 (19.53%)	
Widowed	158 (79.40%)	41 (20.60%)		124 (78.48%)	34 (21.52%)		34 (82.93%)	7 (17.07%)	

TABLE I.2 ALCOHOL CONSUMPTION BELOW OR ABOVE THE WEEKLY LIMIT IN URBAN & RURAL, URBAN AND RURAL DATASETS

Variable	Urban & rural dataset			Urban dataset			Rural dataset		
	Above the limit		p-value[2]	Above the limit		p-value[2]	Above the limit		p-value[2]
	No	Yes		No	Yes		No	Yes	
	N = 8,444	N = 2,669		N = 6,767	N = 2,139		N = 1,677	N = 530	
General health (genhelf2)	<0.001			<0.001			<0.001		
Very good	2,903 (74.38%)	1,000 (25.62%)		2,264 (73.72%)	807 (26.28%)		639 (76.80%)	193 (23.20%)	
Good	3,258 (74.52%)	1,114 (25.48%)		2,617 (75.14%)	866 (24.86%)		641 (72.10%)	248 (27.90%)	
Fair	1,519 (79.24%)	398 (20.76%)		1,244 (79.13%)	328 (20.87%)		275 (79.71%)	70 (20.29%)	
Bad	566 (82.03%)	124 (17.97%)		475 (81.48%)	108 (18.52%)		91 (85.05%)	16 (14.95%)	
Very bad	198 (86.09%)	32 (13.91%)		167 (84.77%)	30 (15.23%)		31 (93.94%)	2 (6.06%)	
LTC limitac_h	<0.001			<0.001			<0.001		
No limitations	4,569 (74.10%)	1,597 (25.90%)		3,640 (74.29%)	1,260 (25.71%)		929 (73.38%)	337 (26.62%)	
Not at all	1,078 (72.89%)	401 (27.11%)		847 (72.52%)	321 (27.48%)		231 (74.28%)	80 (25.72%)	
A little	1,280 (76.74%)	388 (23.26%)		1,024 (76.42%)	316 (23.58%)		256 (78.05%)	72 (21.95%)	
A lot	1,515 (84.26%)	283 (15.74%)		1,255 (83.83%)	242 (16.17%)		260 (86.38%)	41 (13.62%)	
Life satisfaction (lifesat2)			<0.001			<0.001			0.93
Above mode (mode = 8)	2,798 (74.75%)	945 (25.25%)		2,137 (74.28%)	740 (25.72%)		661 (76.33%)	205 (23.67%)	
Mode (mode = 8)	2,588 (74.54%)	884 (25.46%)		2,051 (74.28%)	710 (25.72%)		537 (75.53%)	174 (24.47%)	
Below mode (mode = 8)	3,044 (78.49%)	834 (21.51%)		2,567 (78.96%)	684 (21.04%)		477 (76.08%)	150 (23.92%)	
Activity level (adt10gpTW)			<0.001			<0.001			0.013
Meets recommendations	5,797 (74.01%)	2,036 (25.99%)		4,578 (73.85%)	1,621 (26.15%)		1,219 (74.60%)	415 (25.40%)	
Some activity	851 (77.58%)	246 (22.42%)		699 (77.67%)	201 (22.33%)		152 (77.16%)	45 (22.84%)	
Low activity	305 (77.81%)	87 (22.19%)		249 (78.80%)	67 (21.20%)		56 (73.68%)	20 (26.32%)	
Very low activity	1,445 (83.29%)	290 (16.71%)		1,204 (83.26%)	242 (16.74%)		241 (83.39%)	48 (16.61%)	
Smoking (cig)			<0.001			<0.001			<0.001
Never smoked	4,695 (81.23%)	1,085 (18.77%)		3,769 (81.39%)	862 (18.61%)		926 (80.59%)	223 (19.41%)	
Ex-smoker	2,365 (70.58%)	986 (29.42%)		1,838 (70.13%)	783 (29.87%)		527 (72.19%)	203 (27.81%)	
Light smokers	472 (70.45%)	198 (29.55%)		405 (70.93%)	166 (29.07%)		67 (67.68%)	32 (32.32%)	
Moderate smokers	567 (70.35%)	239 (29.65%)		471 (70.93%)	193 (29.07%)		96 (67.61%)	46 (32.39%)	
Heavy smokers	302 (68.79%)	137 (31.21%)		252 (68.85%)	114 (31.15%)		50 (68.49%)	23 (31.51%)	
Educational qualification (hedqul08)			<0.001			0.017			
Degree or higher	3,595 (74.90%)	1,205 (25.10%)		2,861 (75.21%)	943 (24.79%)		734 (73.69%)	262 (26.31%)	
HNC/D	1,228 (74.88%)	412 (25.12%)		992 (74.92%)	332 (25.08%)		236 (74.68%)	80 (25.32%)	
Higher grade or equivalent	1,181 (75.18%)	390 (24.82%)		916 (74.78%)	309 (25.22%)		265 (76.59%)	81 (23.41%)	
Standard grade or equivalent	1,451 (77.22%)	428 (22.78%)		1,189 (77.06%)	354 (22.94%)		262 (77.98%)	74 (22.02%)	
Other school level	154 (82.80%)	32 (17.20%)		138 (83.13%)	28 (16.87%)		16 (80.00%)	4 (20.00%)	
No qualifications	815 (80.37%)	199 (19.63%)		653 (79.34%)	170 (20.66%)		162 (84.82%)	29 (15.18%)	
Social Class (schrpg7)			<0.001			<0.001			0.014
I Professional	798 (73.75%)	284 (26.25%)		649 (74.43%)	223 (25.57%)		149 (70.95%)	61 (29.05%)	
II Managerial technical	3,200 (73.38%)	1,161 (26.62%)		2,502 (73.24%)	914 (26.76%)		698 (73.86%)	247 (26.14%)	
IIIN Skilled non-manual	1,200 (79.37%)	312 (20.63%)		1,012 (79.37%)	263 (20.63%)		188 (79.32%)	49 (20.68%)	
IIIM Skilled manual	1,434 (76.36%)	444 (23.64%)		1,142 (76.70%)	347 (23.30%)		292 (75.06%)	97 (24.94%)	
IV Semi-skilled manual	1,155 (79.60%)	296 (20.40%)		917 (78.78%)	247 (21.22%)		238 (82.93%)	49 (17.07%)	
V Unskilled manual	360 (77.75%)	103 (22.25%)		287 (77.15%)	85 (22.85%)		73 (80.22%)	18 (19.78%)	

TABLE I.2 ALCOHOL CONSUMPTION BELOW OR ABOVE THE WEEKLY LIMIT IN URBAN & RURAL, URBAN AND RURAL DATASETS

Variable	Urban & rural dataset			Urban dataset			Rural dataset		
	Above the limit		p-value[2]	Above the limit		p-value[2]	Above the limit		p-value[2]
	No	Yes		No	Yes		No	Yes	
	N = 8,444	N = 2,669		N = 6,767	N = 2,139		N = 1,677	N = 530	
Others	187 (88.63%)	24 (11.37%)		167 (89.30%)	20 (10.70%)		20 (83.33%)	4 (16.67%)	
Economic activity (neconacb)			<0.001						
In employment	6,288 (74.76%)	2,123 (25.24%)		4,988 (74.76%)	1,684 (25.24%)		1,300 (74.76%)	439 (25.24%)	
ILO unemployed	198 (75.57%)	64 (24.43%)		170 (75.89%)	54 (24.11%)		28 (73.68%)	10 (26.32%)	
Inactive	1,953 (80.21%)	482 (19.79%)		1,604 (80.00%)	401 (20.00%)		349 (81.16%)	81 (18.84%)	
SIMD quintile (SIMD20_RPa)			<0.001			<0.001			0.054
Least deprived	1,613 (69.77%)	699 (30.23%)		1,462 (69.65%)	637 (30.35%)		151 (70.89%)	62 (29.11%)	
4th	1,825 (74.76%)	616 (25.24%)		1,173 (75.00%)	391 (25.00%)		652 (74.34%)	225 (25.66%)	
3rd	1,781 (77.43%)	519 (22.57%)		1,121 (77.52%)	325 (22.48%)		660 (77.28%)	194 (22.72%)	
2nd	1,726 (78.78%)	465 (21.22%)		1,553 (78.51%)	425 (21.49%)		173 (81.22%)	40 (18.78%)	
Most deprived	1,499 (80.20%)	370 (19.80%)		1,458 (80.15%)	361 (19.85%)		41 (82.00%)	9 (18.00%)	
Urban-rural class			>0.99						
Urban	6,767 (75.98%)	2,139 (24.02%)							
Rural	1,677 (75.99%)	530 (24.01%)							
Total Units of alcohol/week (drating)			<0.001			<0.001			<0.001
Mean (SD)	3.58 (4.06)	30.93 (26.84)		3.54 (4.06)	31.32 (27.60)		3.76 (4.09)	29.34 (23.49)	
Median (IQR)	1.84 (0.09, 6.26)	23.07 (17.91, 34.10)		1.78 (0.07, 6.12)	23.43 (17.85, 34.50)		2.00 (0.17, 6.75)	22.40 (18.00, 33.08)	
Range	0.00 - 14.00	14.01 - 412.50		0.00 - 14.00	14.03 - 412.50		0.00 - 14.00	14.01 - 302.50	
Units of normal beer/week (nberwu)			<0.001			<0.001			<0.001
Mean (SD)	1.01 (2.22)	11.21 (18.40)		1.02 (2.26)	11.64 (19.38)		0.94 (2.06)	9.51 (13.64)	
Median (IQR)	0.00 (0.00, 0.69)	5.08 (0.03, 15.23)		0.00 (0.00, 0.69)	6.00 (0.04, 15.23)		0.00 (0.00, 0.67)	4.35 (0.00, 14.00)	
Range	0.00 - 14.00	0.00 - 280.00		0.00 - 14.00	0.00 - 280.00		0.00 - 14.00	0.00 - 105.00	
Units of strong beer/week (sberwu)			<0.001			<0.001			<0.001
Mean (SD)	0.03 (0.30)	0.98 (8.61)		0.03 (0.29)	1.16 (9.56)		0.03 (0.33)	0.26 (1.81)	
Median (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)		0.00 (0.00, 0.00)	0.00 (0.00, 0.00)		0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	
Range	0.00 - 9.00	0.00 - 210.00		0.00 - 9.00	0.00 - 210.00		0.00 - 6.96	0.00 - 24.00	
Units of spirits/week (spirwu)			<0.001			<0.001			<0.001
Mean (SD)	0.93 (1.81)	6.04 (14.30)		0.94 (1.83)	6.21 (15.11)		0.89 (1.73)	5.33 (10.44)	
Median (IQR)	0.12 (0.00, 0.81)	1.88 (0.23, 7.00)		0.12 (0.00, 0.92)	2.25 (0.23, 7.00)		0.12 (0.00, 0.75)	1.50 (0.23, 6.00)	
Range	0.00 - 14.00	0.00 - 280.00		0.00 - 14.00	0.00 - 280.00		0.00 - 13.50	0.00 - 154.00	
Units of sherry/week (sherwu)			<0.001			<0.001			<0.001
Mean (SD)	0.03 (0.32)	0.27 (2.42)		0.03 (0.29)	0.29 (2.59)		0.05 (0.41)	0.19 (1.53)	
Median (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)		0.00 (0.00, 0.00)	0.00 (0.00, 0.00)		0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	
Range	0.00 - 10.50	0.00 - 56.00		0.00 - 9.00	0.00 - 56.00		0.00 - 10.50	0.00 - 30.00	
Units of wine/week (winewu)			<0.001			<0.001			<0.001
Mean (SD)	1.56 (2.76)	12.29 (14.91)		1.50 (2.71)	11.87 (13.75)		1.83 (2.92)	13.97 (18.79)	
Median (IQR)	0.09 (0.00, 1.69)	9.00 (0.84, 15.75)		0.06 (0.00, 1.50)	9.00 (0.69, 15.75)		0.17 (0.00, 2.25)	12.00 (2.25, 20.81)	
Range	0.00 - 14.00	0.00 - 297.00		0.00 - 14.00	0.00 - 126.00		0.00 - 14.00	0.00 - 297.00	

TABLE I.2 ALCOHOL CONSUMPTION BELOW OR ABOVE THE WEEKLY LIMIT IN URBAN & RURAL, URBAN AND RURAL DATASETS

Variable	Urban & rural dataset			Urban dataset			Rural dataset		
	Above the limit		p-value[2]	Above the limit		p-value[2]	Above the limit		p-value[2]
	No	Yes		No	Yes		No	Yes	
	N = 8,444	N = 2,669		N = 6,767	N = 2,139		N = 1,677	N = 530	
Units of alcopops/week (popswu)			0.53			0.36			0.60
Mean (SD)	0.03 (0.23)	0.14 (1.42)		0.03 (0.25)	0.15 (1.49)		0.01 (0.13)	0.08 (1.05)	
Median (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)		0.00 (0.00, 0.00)	0.00 (0.00, 0.00)		0.0 (0.00, 0.00)	0.00 (0.00, 0.00)	
Range	0.00 - 8.25	0.00 - 36.00		0.00 - 8.25	0.00 - 36.00		0.00 - 3.94	0.00 - 22.50	

[1] n (%)

[2] Pearson's Chi-squared test; Wilcoxon rank sum test

TABLE I.3 COEFICIENTS AFTER CROSS-VALIDATION IN THE URBAN & RURAL TRAINING DATASET (dw.train)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.767	0.125	-6.160	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.688	0.057	-11.996	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.119	0.094	1.272	0.20
45-54	0.470	0.089	5.257	0.00
55-64	0.648	0.091	7.115	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.106	0.083	-1.272	0.20
Other	-0.661	0.122	-5.434	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.014	0.074	-0.187	0.85
Roman Catholic	-0.009	0.089	-0.100	0.92
Other religion	-0.505	0.113	-4.464	0.00
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.047	0.083	-0.566	0.57
A little	-0.273	0.084	-3.269	0.00
A lot	-0.527	0.103	-5.106	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	0.000	0.088	0.005	1.00
Higher grade	-0.118	0.090	-1.305	0.19
Standard School Grade	-0.204	0.089	-2.298	0.02
Not qualifications	-0.391	0.126	-3.103	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.042	0.099	0.423	0.67
IIIN Skilled non-manual	-0.184	0.123	-1.495	0.13
IIIM Skilled manual	-0.156	0.119	-1.308	0.19
IV Semi-skilled manual	-0.246	0.131	-1.871	0.06
V Unskilled manual & other	-0.213	0.166	-1.286	0.20
Employment Status (neconacb)				
In employment				
ILO unemployed & Inactive	-0.151	0.079	-1.902	0.06
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.202	0.083	-2.449	0.01
3rd	-0.325	0.087	-3.715	0.00
2nd	-0.494	0.092	-5.340	0.00
Most deprived	-0.434	0.102	-4.258	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.183	0.087	-2.098	0.04
Very low activity	-0.354	0.096	-3.698	0.00
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.659	0.064	10.233	0.00
Light smokers	0.974	0.121	8.062	0.00
Moderate smokers	0.901	0.111	8.103	0.00
Heavy smokers	0.978	0.146	6.707	0.00

TABLE I.4 COEFICIENTS AFTER CROSS-VALIDATION IN THE URBAN & RURAL OVERSAMPLED TRAINING DATASET - RATIO 30% (up30dw.train)

Variable-category	Estimate	Std. Error	Statistic	p-value
Intercept	-0.767	0.125	-6.160	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.710	0.052	-13.703	0.00
Age (ag16g10)				
25-44	-	-	-	-
35-44	0.083	0.086	0.967	0.33
45-54	0.492	0.083	5.931	0.00
55-64	0.694	0.087	8.010	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.099	0.075	-1.330	0.18
Other	-0.635	0.107	-5.942	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.006	0.066	-0.086	0.93
Roman Catholic	-0.050	0.081	-0.615	0.54
Other religion	-0.436	0.099	-4.390	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	0.017	0.076	0.218	0.83
Single	-0.102	0.077	-1.330	0.18
Separated & Widowed	-0.300	0.089	-3.378	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.131	0.060	2.182	0.03
Fair	-0.060	0.091	-0.656	0.51
Bad_ Very bad	-0.170	0.155	-1.100	0.27
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.103	0.076	-1.350	0.18
A little	-0.394	0.082	-4.830	0.00
A lot	-0.450	0.110	-4.084	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.022	0.080	-0.269	0.79
Higher grade	-0.174	0.082	-2.123	0.03
Standard School Grade	-0.158	0.079	-1.986	0.05
Not qualifications	-0.424	0.113	-3.741	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.051	0.090	0.568	0.57
IIIN Skilled non-manual	-0.174	0.112	-1.556	0.12
IIIM Skilled manual	-0.103	0.107	-0.955	0.34
IV Semi-skilled manual	-0.155	0.118	-1.312	0.19
V Unskilled manual & other	-0.311	0.153	-2.038	0.04
Employment Status (neconacb)				
In employment	-	-	-	-
ILO unemployed & Inactive	-0.122	0.072	-1.705	0.09
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.152	0.076	-2.016	0.04
3rd	-0.229	0.079	-2.884	0.00
2nd	-0.339	0.083	-4.067	0.00
Most deprived	-0.376	0.095	-3.972	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.292	0.081	-3.614	0.00
Very low activity	-0.264	0.085	-3.091	0.00
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.626	0.058	10.742	0.00
Light smokers	0.915	0.111	8.226	0.00
Moderate smokers	0.887	0.101	8.747	0.00
Heavy smokers	1.152	0.129	8.921	0.00

TABLE I.1 COEFICIENTS AFTER CROSS-VALIDATION IN THE URBAN & RURAL OVERSAMPLED TRAINING DATASET - RATIO 40% (up40dw.train)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.138	0.101	-1.375	0.17
Sex (sex)				
Male	-	-	-	-
Female	-0.709	0.045	-15.693	0.00
Age (ag16g10)				
25-44	-	-	-	-
35-44	0.200	0.073	2.754	0.01
45-54	0.558	0.070	7.986	0.00
55-64	0.680	0.072	9.479	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.134	0.066	-2.024	0.04
Other	-0.656	0.091	-7.236	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.059	0.058	-1.016	0.31
Roman Catholic	0.031	0.069	0.454	0.65
Other religion	-0.516	0.087	-5.962	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.039	0.053	0.745	0.46
Fair	-0.174	0.079	-2.203	0.03
Bad_ Very bad	-0.061	0.129	-0.473	0.64
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.031	0.067	-0.464	0.64
A little	-0.201	0.069	-2.912	0.00
A lot	-0.486	0.097	-5.014	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.021	0.070	-0.300	0.76
Higher grade	-0.162	0.072	-2.241	0.03
Standard School Grade	-0.092	0.069	-1.332	0.18
Not qualifications	-0.350	0.097	-3.594	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.098	0.079	1.245	0.21
IIIN Skilled non-manual	-0.080	0.097	-0.828	0.41
IIIM Skilled manual	-0.136	0.095	-1.429	0.15
IV Semi-skilled manual	-0.245	0.104	-2.360	0.02
V Unskilled manual & other	-0.220	0.130	-1.700	0.09
Employment Status (neconacb)				
In employment	-	-	-	-
ILO unemployed & Inactive	-0.090	0.062	-1.450	0.15
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.147	0.068	-2.159	0.03
3rd	-0.260	0.072	-3.629	0.00
2nd	-0.344	0.072	-4.802	0.00
Most deprived	-0.452	0.081	-5.598	0.00
Urban-rural class				
Urban	-	-	-	-
Rural	-0.131	0.060	-2.172	0.03
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.260	0.069	-3.764	0.00
Very low activity	-0.371	0.075	-4.921	0.00
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.671	0.051	13.257	0.00
Light smokers	0.992	0.096	10.327	0.00
Moderate smokers	0.825	0.089	9.288	0.00
Heavy smokers	1.085	0.114	9.530	0.00

TABLE I.2 COEFICIENTS AFTER CROSS-VALIDATION IN THE URBAN & RURAL OVERSAMPLED TRAINING DATASET – RATIO 50% (up50dw.train)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	0.276	0.094	2.943	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.709	0.041	-17.406	0.00
Age (ag16g10)				
25-44	-	-	-	-
35-44	0.122	0.066	1.850	0.06
45-54	0.489	0.065	7.555	0.00
55-64	0.652	0.068	9.628	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.082	0.059	-1.387	0.17
Other	-0.678	0.080	-8.439	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	0.045	0.052	0.866	0.39
Roman Catholic	-0.009	0.062	-0.143	0.89
Other religion	-0.511	0.077	-6.644	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	0.101	0.059	1.697	0.09
Single	-0.065	0.060	-1.088	0.28
Separated & Widowed	-0.174	0.067	-2.585	0.01
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	0.028	0.059	0.471	0.64
A little	-0.295	0.058	-5.052	0.00
A lot	-0.523	0.071	-7.353	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	0.069	0.063	1.104	0.27
Higher grade	-0.029	0.064	-0.452	0.65
Standard School Grade	-0.129	0.062	-2.075	0.04
Not qualifications	-0.290	0.087	-3.318	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	-0.079	0.088	-0.905	0.37
IIIN Skilled non-manual	-0.125	0.086	-1.459	0.14
IIIM Skilled manual	-0.146	0.093	-1.566	0.12
IV Semi-skilled manual	-0.147	0.115	-1.275	0.20
V Unskilled manual & other	-0.097	0.055	-1.759	0.08
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.153	0.061	-2.490	0.01
3rd	-0.297	0.065	-4.591	0.00
2nd	-0.468	0.066	-7.129	0.00
Most deprived	-0.455	0.072	-6.285	0.00
Urban-rural class				
Urban				
Rural	-0.126	0.054	-2.326	0.02
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.159	0.060	-2.661	0.01
Very low activity	-0.470	0.067	-7.014	0.00
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.625	0.046	13.717	0.00
Light smokers	0.976	0.088	11.048	0.00
Moderate smokers	0.898	0.079	11.347	0.00
Heavy smokers	0.997	0.105	9.508	0.00

TABLE I.7 COEFICIENTS AFTER CROSS-VALIDATION IN THE SMOTE URBAN & RURAL TRAINING DATASET k = 3 / or = 0.5 (smk3r50_dw)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.383	0.126	-3.036	0.00
Survey year (syear)				
2017	-	-	-	-
2018	0.167	0.075	2.221	0.03
2019	0.250	0.074	3.368	0.00
2021	0.357	0.075	4.740	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.740	0.050	-14.784	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.097	0.084	1.166	0.24
45-54	0.560	0.081	6.934	0.00
55-64	0.764	0.085	9.027	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.214	0.074	-2.904	0.00
Other	-0.831	0.108	-7.720	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.099	0.065	-1.532	0.13
Roman Catholic	-0.146	0.079	-1.839	0.07
Other religion	-0.658	0.101	-6.520	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	-0.042	0.074	-0.561	0.57
Single	-0.093	0.074	-1.250	0.21
Separated & Widowed	-0.290	0.086	-3.362	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.126	0.058	2.189	0.03
Fair	-0.106	0.089	-1.189	0.23
Bad_ Very bad	0.153	0.150	1.024	0.31
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.211	0.075	-2.817	0.00
A little	-0.365	0.078	-4.688	0.00
A lot	-0.554	0.112	-4.951	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.123	0.078	-1.569	0.12
Higher grade	-0.209	0.079	-2.632	0.01
Standard School Grade	-0.282	0.078	-3.601	0.00
Not qualifications	-0.511	0.114	-4.473	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.109	0.086	1.269	0.20
IIIN Skilled non-manual	-0.177	0.107	-1.650	0.10
IIIM Skilled manual	-0.101	0.104	-0.970	0.33
IV Semi-skilled manual	-0.218	0.116	-1.873	0.06
V Unskilled manual & other	-0.242	0.149	-1.623	0.10
Employment Status (neconacb)				
In employment	-	-	-	-
ILO unemployed & Inactive	-0.183	0.071	-2.582	0.01
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.173	0.074	-2.333	0.02
3rd	-0.226	0.078	-2.891	0.00
2nd	-0.442	0.080	-5.501	0.00
Most deprived	-0.423	0.091	-4.654	0.00
Urban-rural class				
Urban	-	-	-	-
Rural	-0.229	0.067	-3.398	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.372	0.079	-4.694	0.00
Very low activity	-0.472	0.087	-5.395	0.00

Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.674	0.056	12.055	0.00
Light smokers	0.915	0.111	8.253	0.00
Moderate smokers	0.866	0.101	8.584	0.00
Heavy smokers	0.960	0.135	7.084	0.00

TABLE I.8 COEFICIENTS AFTER CROSS-VALIDATION IN THE SMOTE URBAN & RURAL TRAINING DATASET k = 3 / or = 0.75 (smk3r75_dw)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.101	0.114	-0.886	0.38
Survey year (syear)				
2017	-	-	-	-
2018	0.210	0.069	3.040	0.00
2019	0.393	0.067	5.842	0.00
2021	0.546	0.068	8.018	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.734	0.045	-16.310	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.136	0.074	1.829	0.07
45-54	0.575	0.073	7.919	0.00
55-64	0.755	0.077	9.818	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.279	0.067	-4.140	0.00
Other	-0.904	0.096	-9.388	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.109	0.058	-1.858	0.06
Roman Catholic	-0.203	0.072	-2.830	0.00
Other religion	-0.767	0.091	-8.435	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	-0.087	0.068	-1.295	0.20
Single	-0.107	0.068	-1.585	0.11
Separated & Widowed	-0.231	0.077	-3.003	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.111	0.052	2.135	0.03
Fair	-0.144	0.082	-1.764	0.08
Bad_ Very bad	0.344	0.137	2.515	0.01
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.222	0.067	-3.325	0.00
A little	-0.454	0.071	-6.403	0.00
A lot	-0.718	0.104	-6.895	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.148	0.071	-2.087	0.04
Higher grade	-0.248	0.072	-3.466	0.00
Standard School Grade	-0.237	0.070	-3.366	0.00
Not qualifications	-0.437	0.103	-4.256	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.282	0.078	3.626	0.00
IIIN Skilled non-manual	-0.136	0.097	-1.395	0.16
IIIM Skilled manual	-0.031	0.095	-0.331	0.74
IV Semi-skilled manual	-0.193	0.106	-1.823	0.07
V Unskilled manual & other	-0.244	0.136	-1.801	0.07
Employment Status (neconacb)				
In employment	-	-	-	-
ILO unemployed & Inactive	-0.236	0.065	-3.634	0.00
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.171	0.067	-2.568	0.01
3rd	-0.284	0.071	-4.019	0.00
2nd	-0.485	0.073	-6.655	0.00
Most deprived	-0.378	0.082	-4.629	0.00
urbrur_allrural	-0.241	0.061	-3.959	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-0.472	0.072	-6.547	0.00
Low activity	-0.469	0.078	-5.988	0.00
Very low activity				
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.664	0.050	13.266	0.00
Light smokers	0.911	0.101	9.040	0.00
Moderate smokers	0.802	0.093	8.599	0.00
Heavy smokers	0.848	0.125	6.813	0.00

TABLE I.9 COEFICIENTS AFTER CROSS-VALIDATION IN THE SMOTE URBAN & RURAL TRAINING DATASET k = 3 / or = 1 (smk3r1_dw)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	0.164	0.108	1.522	0.13
Survey year (syear)				
2017	-	-	-	-
2018	0.253	0.065	3.893	0.00
2019	0.463	0.063	7.352	0.00
2021	0.643	0.064	10.063	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.727	0.042	-17.241	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.164	0.070	2.350	0.02
45-54	0.633	0.068	9.262	0.00
55-64	0.870	0.072	12.071	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.292	0.063	-4.639	0.00
Other	-0.982	0.090	-10.955	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.141	0.055	-2.576	0.01
Roman Catholic	-0.294	0.068	-4.361	0.00
Other religion	-0.843	0.085	-9.896	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	-0.094	0.063	-1.492	0.14
Single	-0.145	0.063	-2.291	0.02
Separated & Widowed	-0.329	0.073	-4.529	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.196	0.049	4.034	0.00
Fair	-0.076	0.077	-0.992	0.32
Bad_ Very bad	0.482	0.128	3.755	0.00
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.325	0.063	-5.127	0.00
A little	-0.570	0.067	-8.518	0.00
A lot	-0.746	0.097	-7.689	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.264	0.067	-3.943	0.00
Higher grade	-0.284	0.067	-4.247	0.00
Standard School Grade	-0.315	0.066	-4.761	0.00
Not qualifications	-0.453	0.095	-4.765	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.234	0.073	3.226	0.00
IIIN Skilled non-manual	-0.185	0.091	-2.042	0.04
IIIM Skilled manual	-0.028	0.089	-0.316	0.75
IV Semi-skilled manual	-0.141	0.097	-1.451	0.15
V Unskilled manual & other	-0.269	0.127	-2.122	0.03
Employment Status (neconacb)				
In employment	-	-	-	-
ILO unemployed & Inactive	-0.257	0.061	-4.211	0.00
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.176	0.063	-2.810	0.00
3rd	-0.263	0.066	-3.970	0.00
2nd	-0.508	0.068	-7.441	0.00
Most deprived	-0.321	0.076	-4.210	0.00
urbrur_allrural	-0.264	0.057	-4.638	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.532	0.067	-7.892	0.00
Very low activity	-0.532	0.074	-7.200	0.00
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.655	0.047	13.947	0.00
Light smokers	0.881	0.095	9.267	0.00
Moderate smokers	0.684	0.089	7.717	0.00
Heavy smokers	0.846	0.118	7.147	0.00

TABLE I.10 COEFICIENTS AFTER CROSS-VALIDATION IN THE SMOTE URBAN & RURAL TRAINING DATASET k = 5 / or = 0.5 (smk5r50_dw)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.452	0.127	-3.556	0.00
Survey year (syear)				
2017	-	-	-	-
2018	0.126	0.076	1.667	0.10
2019	0.276	0.074	3.738	0.00
2021	0.377	0.075	5.023	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.687	0.050	-13.732	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.110	0.084	1.316	0.19
45-54	0.523	0.081	6.438	0.00
55-64	0.767	0.085	9.031	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.226	0.074	-3.051	0.00
Other	-0.853	0.109	-7.794	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.093	0.065	-1.443	0.15
Roman Catholic	-0.132	0.079	-1.667	0.10
Other religion	-0.670	0.102	-6.593	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	-0.073	0.075	-0.982	0.33
Single	-0.185	0.076	-2.437	0.01
Separated & Widowed	-0.262	0.085	-3.092	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.117	0.058	2.024	0.04
Fair	-0.130	0.090	-1.445	0.15
Bad_ Very bad	0.146	0.148	0.984	0.33
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.094	0.073	-1.282	0.20
A little	-0.377	0.079	-4.796	0.00
A lot	-0.506	0.111	-4.545	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.047	0.077	-0.613	0.54
Higher grade	-0.214	0.080	-2.684	0.01
Standard School Grade	-0.259	0.079	-3.295	0.00
Not qualifications	-0.408	0.112	-3.660	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.163	0.086	1.886	0.06
IIIN Skilled non-manual	-0.172	0.108	-1.586	0.11
IIIM Skilled manual	-0.139	0.105	-1.318	0.19
IV Semi-skilled manual	-0.127	0.115	-1.099	0.27
V Unskilled manual & other	-0.234	0.150	-1.563	0.12
Employment Status (neconacb)				
In employment	-	-	-	-
ILO unemployed & Inactive	-0.177	0.071	-2.514	0.01
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.169	0.074	-2.289	0.02
3rd	-0.292	0.079	-3.695	0.00
2nd	-0.484	0.081	-5.969	0.00
Most deprived	-0.356	0.090	-3.962	0.00
Urban-rural class				
Urban	-	-	-	-
Rural	-0.180	0.067	-2.678	0.01
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.384	0.079	-4.842	0.00
Very low activity	-0.482	0.087	-5.539	0.00

Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.681	0.056	12.218	0.00
Light smokers	0.914	0.111	8.225	0.00
Moderate smokers	0.880	0.101	8.685	0.00
Heavy smokers	0.936	0.135	6.911	0.00

TABLE I.11 COEFICIENTS AFTER CROSS-VALIDATION IN THE SMOTE URBAN & RURAL TRAINING DATASET k = 5 / or = 0. 75 (smk5r75_dw)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.062	0.115	-0.538	0.59
Survey year (syear)				
2017	-	-	-	-
2018	0.218	0.069	3.146	0.00
2019	0.396	0.068	5.867	0.00
2021	0.623	0.068	9.137	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.717	0.045	-15.928	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.088	0.074	1.188	0.23
45-54	0.507	0.072	7.007	0.00
55-64	0.714	0.076	9.378	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.309	0.067	-4.596	0.00
Other	-0.977	0.097	-10.037	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.134	0.059	-2.291	0.02
Roman Catholic	-0.158	0.071	-2.221	0.03
Other religion	-0.807	0.093	-8.698	0.00
Marital status (maritalg)				
Married / civil partnership				
Living as married	-0.046	0.067	-0.688	0.49
Single	-0.138	0.068	-2.041	0.04
Separated & Widowed	-0.299	0.078	-3.806	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.184	0.052	3.543	0.00
Fair	-0.115	0.083	-1.395	0.16
Bad_ Very bad	0.245	0.135	1.815	0.07
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.276	0.068	-4.079	0.00
A little	-0.521	0.071	-7.308	0.00
A lot	-0.579	0.101	-5.715	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.146	0.071	-2.062	0.04
Higher grade	-0.237	0.072	-3.301	0.00
Standard School Grade	-0.265	0.071	-3.735	0.00
Not qualifications	-0.302	0.100	-3.031	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.250	0.078	3.218	0.00
IIIN Skilled non-manual	-0.171	0.098	-1.749	0.08
IIIM Skilled manual	-0.028	0.095	-0.295	0.77
IV Semi-skilled manual	-0.108	0.105	-1.035	0.30
V Unskilled manual & other	-0.247	0.136	-1.825	0.07
Employment Status (neconacb)				
In employment	-	-	-	-
ILO unemployed & Inactive	-0.203	0.064	-3.142	0.00
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.190	0.066	-2.868	0.00
3rd	-0.333	0.071	-4.673	0.00
2nd	-0.575	0.073	-7.873	0.00
Most deprived	-0.424	0.082	-5.166	0.00
urbrur_allrural	-0.238	0.061	-3.921	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.445	0.072	-6.223	0.00
Very low activity	-0.510	0.079	-6.464	0.00

Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.654	0.050	13.043	0.00
Light smokers	0.797	0.103	7.745	0.00
Moderate smokers	0.827	0.093	8.910	0.00
Heavy smokers	0.903	0.123	7.351	0.00

TABLE I.12 COEFICIENTS AFTER CROSS-VALIDATION IN THE SMOTE URBAN & RURAL TRAINING DATASET k = 5 / or = 1 (smk5r100_dw)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	0.043	0.108	0.398	0.69
Survey year (syear)				
2017	-	-	-	-
2018	0.265	0.066	4.022	0.00
2019	0.529	0.064	8.305	0.00
2021	0.722	0.064	11.209	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.782	0.043	-18.388	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.180	0.070	2.563	0.01
45-54	0.619	0.069	8.963	0.00
55-64	0.801	0.073	11.009	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.251	0.107	-2.338	0.02
Other	-0.204	0.145	-1.400	0.16
Birthplace (birthpla3)				
Scotland	-	-	-	-
Rest uk	0.051	0.105	-0.481	0.63
Elsewhere	0.902	0.153	-5.879	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.034	0.055	-0.614	0.54
Roman Catholic	-0.128	0.067	-1.909	0.06
Other religion	-0.884	0.088	-10.010	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	-0.122	0.064	-1.894	0.06
Single	-0.178	0.064	-2.782	0.01
Separated & Widowed	-0.267	0.073	-3.666	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.153	0.049	3.136	0.00
Fair	-0.088	0.077	-1.141	0.25
Bad_ Very bad	0.417	0.129	3.242	0.00
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.299	0.063	-4.708	0.00
A little	-0.578	0.067	-8.603	0.00
A lot	-0.810	0.098	-8.232	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.223	0.067	-3.309	0.00
Higher grade	-0.360	0.069	-5.245	0.00
Standard School Grade	-0.271	0.067	-4.058	0.00
Not qualifications	-0.397	0.095	-4.167	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.350	0.074	4.752	0.00
IIIN Skilled non-manual	-0.189	0.093	-2.035	0.04
IIIM Skilled manual	-0.047	0.090	-0.523	0.60
IV Semi-skilled manual	-0.167	0.100	-1.676	0.09
V Unskilled manual & other	-0.335	0.129	-2.593	0.01
Employment Status (neconacb)				
In employment	-	-	-	-
ILO unemployed & Inactive	-0.203	0.062	-3.294	0.00
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.227	0.064	-3.564	0.00
3rd	-0.290	0.067	-4.301	0.00
2nd	-0.428	0.068	-6.274	0.00
Most deprived	-0.340	0.077	-4.413	0.00
urbrur_allrural	-0.273	0.058	-4.725	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.428	0.067	-6.413	0.00
Very low activity	-0.456	0.074	-6.170	0.00

Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.721	0.047	15.239	0.00
Light smokers	0.901	0.096	9.364	0.00
Moderate smokers	0.891	0.088	10.172	0.00
Heavy smokers	0.926	0.118	7.854	0.00

TABLE I.13 COEFICIENTS AFTER CROSS-VALIDATION IN THE SMOTE URBAN & RURAL TRAINING DATASET k = 10 / or = 0. 5 (smk10r50_dw)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.452	0.127	-3.557	0.00
Survey year (syear)				
2017	-	-	-	-
2018	0.158	0.076	2.076	0.04
2019	0.269	0.075	3.591	0.00
2021	0.454	0.075	6.032	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.745	0.050	-14.869	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.172	0.083	2.062	0.04
45-54	0.585	0.081	7.218	0.00
55-64	0.725	0.085	8.519	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.204	0.074	-2.773	0.01
Other	-0.840	0.108	-7.809	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.091	0.065	-1.407	0.16
Roman Catholic	-0.143	0.079	-1.805	0.07
Other religion	-0.673	0.101	-6.646	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	-0.031	0.074	-0.415	0.68
Single	-0.111	0.075	-1.491	0.14
Separated & Widowed	-0.271	0.086	-3.131	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.095	0.058	1.643	0.10
Fair	-0.121	0.090	-1.344	0.18
Bad_ Very bad	0.132	0.150	0.877	0.38
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.184	0.075	-2.474	0.01
A little	-0.365	0.078	-4.672	0.00
A lot	-0.577	0.113	-5.119	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.167	0.078	-2.123	0.03
Higher grade	-0.269	0.080	-3.369	0.00
Standard School Grade	-0.292	0.078	-3.733	0.00
Not qualifications	-0.447	0.112	-3.983	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.204	0.086	2.376	0.02
IIIN Skilled non-manual	-0.117	0.108	-1.082	0.28
IIIM Skilled manual	-0.064	0.105	-0.606	0.54
IV Semi-skilled manual	-0.097	0.116	-0.834	0.40
V Unskilled manual & other	-0.160	0.149	-1.071	0.28
Employment Status (neconacb)				
In employment	-	-	-	-
ILO unemployed & Inactive	-0.180	0.071	-2.525	0.01
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.155	0.073	-2.113	0.03
3rd	-0.330	0.079	-4.159	0.00
2nd	-0.509	0.081	-6.281	0.00
Most deprived	-0.395	0.090	-4.382	0.00
Classification Urban-Rural				
Urban	-	-	-	-
Rural	-0.221	0.068	-3.270	
Physical activity level (adt10gptw)				0.00
Meet recommendations	-	-	-	-
Low activity	-0.319	0.079	-4.058	0.00
Very low activity	-0.425	0.087	-4.905	0.00

Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.662	0.056	11.866	0.00
Light smokers	0.661	0.111	8.010	0.00
Moderate smokers	0.832	0.102	8.166	0.00
Heavy smokers	0.928	0.135	6.859	0.00

TABLE I.14 COEFICIENTS AFTER CROSS-VALIDATION IN THE SMOTE URBAN & RURAL TRAINING DATASET k = 10 / or = 0. 75 (smk10r75_dw)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.068	0.116	-0.590	0.56
Survey year (syear)				
2017	-	-	-	-
2018	0.246	0.070	3.512	0.00
2019	0.416	0.068	6.121	0.00
2021	0.642	0.069	9.347	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.791	0.045	-17.533	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.061	0.074	0.819	0.41
45-54	0.457	0.072	6.307	0.00
55-64	0.667	0.076	8.781	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.338	0.068	-4.969	0.00
Other	-0.971	0.097	-9.989	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.086	0.059	-1.473	0.14
Roman Catholic	-0.200	0.072	-2.774	0.01
Other religion	-0.756	0.092	-8.259	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	-0.169	0.068	-2.500	0.01
Single	-0.203	0.068	-2.979	0.00
Separated & Widowed	-0.309	0.078	-3.961	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.144	0.052	2.770	0.01
Fair	-0.143	0.082	-1.739	0.08
Bad_ Very bad	0.211	0.135	1.559	0.12
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.203	0.067	-3.004	0.00
A little	-0.463	0.071	-6.522	0.00
A lot	-0.583	0.102	-5.696	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.188	0.071	-2.660	0.01
Higher grade	-0.280	0.071	-3.918	0.00
Standard School Grade	-0.297	0.071	-4.209	0.00
Not qualifications	-0.423	0.102	-4.133	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.374	0.079	4.723	0.00
IIIN Skilled non-manual	0.003	0.098	0.029	0.98
IIIM Skilled manual	0.059	0.096	0.613	0.54
IV Semi-skilled manual	-0.055	0.106	-0.521	0.60
V Unskilled manual & other	-0.052	0.135	-0.387	0.70
Employment Status (neconacb)				
In employment	-	-	-	-
ILO unemployed & Inactive	-0.231	0.065	-3.540	0.00
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.227	0.067	-3.410	0.00
3rd	-0.315	0.071	-4.427	0.00
2nd	-0.527	0.073	-7.244	0.00
Most deprived	-0.421	0.081	-5.179	0.00
urbrur_allrural	-0.268	0.061	-4.357	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.416	0.071	-5.824	0.00
Very low activity	-0.453	0.078	-5.810	0.00
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.653	0.050	12.989	0.00
Light smokers	0.797	0.103	7.744	0.00
Moderate smokers	0.827	0.094	8.819	0.00
Heavy smokers	0.954	0.123	7.741	0.00

TABLE I.15 COEFICIENTS AFTER CROSS-VALIDATION IN THE SMOTE URBAN & RURAL TRAINING DATASET k = 10 / or = 1 (smk10r100_dw)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	0.153	0.109	1.408	0.16
Survey year (syear)				
2017	-	-	-	-
2018	0.272	0.066	4.105	0.00
2019	0.485	0.064	7.539	0.00
2021	0.762	0.065	11.745	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.784	0.042	-18.473	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.079	0.070	1.123	0.26
45-54	0.567	0.068	8.278	0.00
55-64	0.800	0.072	11.142	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.328	0.064	-5.147	0.00
Other	-0.978	0.090	-10.816	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.097	0.055	-1.777	0.08
Roman Catholic	-0.255	0.068	-3.749	0.00
Other religion	-0.920	0.088	-10.422	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	-0.072	0.063	-1.131	0.26
Single	-0.178	0.064	-2.765	0.01
Separated & Widowed	-0.323	0.073	-4.403	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.111	0.049	2.259	0.02
Fair	-0.131	0.077	-1.695	0.09
Bad_ Very bad	0.275	0.128	2.154	0.03
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.226	0.064	-3.549	0.00
A little	-0.458	0.067	-6.877	0.00
A lot	-0.617	0.095	-6.468	0.00
Education (hedqul08)				
Degree				
HNC/D	-0.196	0.067	-2.940	0.00
Higher grade	-0.333	0.068	-4.872	0.00
Standard School Grade	-0.319	0.067	-4.738	0.00
Not qualifications	-0.321	0.095	-3.374	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.391	0.074	5.297	0.00
IIIN Skilled non-manual	-0.140	0.093	-1.509	0.13
IIIM Skilled manual	0.004	0.091	0.047	0.96
IV Semi-skilled manual	-0.074	0.100	-0.742	0.46
V Unskilled manual & other	-0.015	0.127	-0.122	0.90
Employment Status (neconacb)				
In employment	-	-	-	-
ILO unemployed & Inactive	-0.285	0.062	-4.606	0.00
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.256	0.063	-4.079	0.00
3rd	-0.379	0.067	-5.653	0.00
2nd	-0.527	0.068	-7.704	0.00
Most deprived	-0.448	0.077	-5.832	0.00
Urban-rural class				
Urban	-	-	-	-
Rural	-0.312	0.058	-5.388	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.485	0.068	-7.181	0.00
Very low activity	-0.542	0.075	-7.271	0.00

Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.683	0.047	14.437	0.00
Light smokers	0.718	0.099	7.224	0.00
Moderate smokers	0.694	0.090	7.748	0.00
Heavy smokers	0.939	0.117	8.047	0.00

TABLE I.16 COEFICIENTS AFTER CROSS-VALIDATION IN THE TRAINING URBAN DATASET (dwu.train)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.802	0.119	-6.749	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.806	0.064	-12.587	0.00
Age (ag16g10)				
25-44	-	-	-	-
35-44	0.105	0.105	1.001	0.32
45-54	0.511	0.099	5.169	0.00
55-64	0.686	0.100	6.873	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.127	0.099	-1.283	0.20
Other	-0.709	0.133	-5.328	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	0.027	0.083	0.322	0.75
Roman Catholic	0.002	0.096	0.022	0.98
Other religion	-0.424	0.129	-3.296	0.00
Life satisfaction (lifesat2)				
Above mode				
Mode	0.028	0.078	0.354	0.72
Below mode	-0.152	0.082	-1.862	0.06
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.021	0.094	-0.221	0.83
A little	-0.173	0.094	-1.854	0.06
A lot	-0.486	0.112	-4.350	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.009	0.097	-0.093	0.93
Higher grade	-0.112	0.100	-1.120	0.26
Standard School Grade	-0.251	0.095	-2.652	0.01
Not qualifications	-0.439	0.131	-3.346	0.00
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.236	0.097	-2.441	0.01
3rd	-0.299	0.102	-2.932	0.00
2nd	-0.483	0.098	-4.942	0.00
Most deprived	-0.476	0.105	-4.524	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.167	0.097	-1.730	0.08
Very low activity	-0.361	0.104	-3.479	0.00
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.716	0.073	9.823	0.00
Light smokers	0.972	0.132	7.386	0.00
Moderate smokers	0.853	0.123	6.919	0.00
Heavy smokers	1.063	0.159	6.697	0.00

TABLE I.17 COEFICIENTS FROM CROSS-VALIDATION IN THE OVERSAMPLING URBAN TRAINING DATASET 30% (up30_dwu)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.556	0.136	-4.071	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.839	0.058	-14.372	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.274	0.096	2.851	0.00
45-54	0.629	0.093	6.734	0.00
55-64	0.794	0.098	8.115	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.082	0.088	-0.926	0.35
Other	-0.548	0.113	-4.848	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	0.035	0.076	0.459	0.65
Roman Catholic	-0.022	0.088	-0.255	0.80
Other religion	-0.303	0.111	-2.727	0.01
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	0.337	0.085	3.958	0.00
Single	0.010	0.086	0.115	0.91
Separated & Widowed	0.116	0.095	1.213	0.23
Life satisfaction (lifesat2)				
Above mode	-	-	-	-
Mode	0.049	0.071	0.694	0.49
Below mode	-0.142	0.076	-1.871	0.06
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	0.038	0.085	0.444	0.66
A little	-0.133	0.085	-1.571	0.12
A lot	-0.423	0.101	-4.191	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	0.146	0.089	1.638	0.10
Higher grade	-0.048	0.094	-0.506	0.61
Standard School Grade	-0.112	0.090	-1.240	0.22
Not qualifications	-0.326	0.126	-2.595	0.01
Social class (schrpg7)				
I Professional				
II Managerial technical	-0.138	0.100	-1.388	0.17
IIIN Skilled non-manual	-0.373	0.124	-3.017	0.00
IIIM Skilled manual	-0.362	0.122	-2.971	0.00
IV Semi-skilled manual	-0.308	0.132	-2.337	0.02
V Unskilled manual & other	-0.211	0.162	-1.299	0.19
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.239	0.088	-2.711	0.01
3rd	-0.272	0.093	-2.931	0.00
2nd	-0.389	0.088	-4.411	0.00
Most deprived	-0.502	0.098	-5.115	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.249	0.089	-2.802	0.01
Very low activity	-0.413	0.094	-4.381	0.00
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.689	0.066	10.417	0.00
Light smokers	0.926	0.120	7.689	0.00
Moderate smokers	0.863	0.112	7.723	0.00
Heavy smokers	1.016	0.146	6.944	0.00

TABLE I.18 COEFICIENTS FROM CROSS-VALIDATION IN THE OVERSAMPLING URBAN TRAINING DATASET 40% (up40_dwu)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.057	0.095	-0.600	0.55
Sex (sex)				
Male	-	-	-	-
Female	-0.841	0.051	-16.631	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.150	0.081	1.854	0.06
45-54	0.492	0.078	6.318	0.00
55-64	0.717	0.079	9.100	0.00
Birthplace (birthpla3)				
Scotland	-	-	-	-
Rest UK	0.290	0.124	2.338	0.02
Elsewhere	0.297	0.146	2.031	0.04
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.360	0.121	-2.978	0.00
Other	-0.852	0.156	-5.454	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	0.083	0.066	1.266	0.21
Roman Catholic	0.025	0.075	0.333	0.74
Other religion	-0.425	0.097	-4.362	0.00
Life satisfaction (lifesat2)				
Above mode	-	-	-	-
Mode	0.005	0.062	0.073	0.94
Below mode	-0.247	0.067	-3.713	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	-0.109	0.061	-1.791	0.07
Fair	-0.240	0.089	-2.701	0.01
Bad_ Very bad	-0.005	0.139	-0.038	0.97
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	0.126	0.076	1.664	0.10
A little	-0.046	0.078	-0.593	0.55
A lot	-0.295	0.105	-2.800	0.01
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	0.001	0.077	0.014	0.99
Higher grade	-0.095	0.079	-1.199	0.23
Standard School Grade	-0.209	0.075	-2.775	0.01
Not qualifications	-0.522	0.103	-5.060	0.00
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.253	0.077	-3.268	0.00
3rd	-0.364	0.082	-4.437	0.00
2nd	-0.479	0.078	-6.159	0.00
Most deprived	-0.384	0.082	-4.655	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.220	0.076	-2.881	0.00
Very low activity	-0.460	0.084	-5.462	0.00
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.749	0.058	12.936	0.00
Light smokers	1.135	0.103	11.049	0.00
Moderate smokers	0.871	0.099	8.803	0.00
Heavy smokers	1.111	0.128	8.677	0.00

TABLE I.19 COEFICIENTS FROM CROSS-VALIDATION IN THE OVERSAMPLING URBAN TRAINING DATASET 50% (up50_dwu)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	0.198	0.092	2.162	0.03
Sex (sex)				
Male	-	-	-	
Female	-0.840	0.046	-18.422	0.00
Age (ag16g10)				
25-34	-	-	-	
35-44	0.241	0.074	3.260	0.00
45-54	0.613	0.073	8.438	0.00
55-64	0.779	0.076	10.253	0.00
Birthplace (birthpla3)				
Scotland	-	-	-	
Rest UK	0.285	0.111	2.568	0.01
Elsewhere	0.195	0.137	1.420	0.16
Ethnic (ethnic05)				
White: Scotland	-	-	-	
White: Rest UK	-0.340	0.108	-3.147	0.00
Other	-0.957	0.145	-6.610	0.00
Religion (religi04)				
None	-	-	-	
Church of Scotland	0.059	1.441	0.15	0.15
Roman Catholic	0.068	-0.379	0.70	0.70
Other religion	0.086	-4.593	0.00	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	
Living as married	0.308	0.068	4.542	0.00
Single	-0.023	0.067	-0.347	0.73
Separated & Widowed	0.108	0.075	1.443	0.15
Life satisfaction (lifesat2)				
Above mode	-	-	-	
Mode	0.042	0.056	0.746	0.46
Below mode	-0.112	0.059	-1.888	0.06
LTC (limitac_h)				
Not limitation	-	-	-	
Not at all	0.104	0.067	1.548	0.12
A little	-0.154	0.066	-2.333	0.02
A lot	-0.536	0.079	-6.794	0.00
Education (hedqul08)				
Degree	-	-	-	
HNC/D	-0.020	0.068	-0.286	0.77
Higher grade	-0.111	0.071	-1.567	0.12
Standard School Grade	-0.253	0.067	-3.777	0.00
Not qualifications	-0.433	0.093	-4.670	0.00
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	
4th	-0.202	0.070	-2.907	0.00
3rd	-0.381	0.074	-5.165	0.00
2nd	-0.469	0.070	-6.706	0.00
Most deprived	-0.401	0.075	-5.315	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	
Low activity	-0.275	0.069	-4.011	0.00
Very low activity	-0.427	0.072	-5.911	0.00
Smoking (cig)				
Never smoked	-	-	-	
Ex-smoker	0.659	0.052	12.706	0.00
Light smokers	0.997	0.095	10.510	0.00
Moderate smokers	0.831	0.088	9.464	0.00
Heavy smokers	1.075	0.115	9.314	0.00

TABLE I.20 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE URBAN TRAINING DATASET K = 3 OR = 0.5 (smk3r50_dwu)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.304	0.118	-2.581	0.01
Year				
2017	-	-	-	-
2018	-0.006	0.083	-0.070	0.94
2019	0.210	0.081	2.575	0.01
2021	0.251	0.083	3.012	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.829	0.056	-14.942	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.149	0.090	1.651	0.10
45-54	0.570	0.086	6.627	0.00
55-64	0.808	0.087	9.295	0.00
Birthplace (birthpla3)				
Scotland	-	-	-	-
Rest UK	0.197	0.141	1.396	0.16
Elsewhere	-0.155	0.178	-0.874	0.38
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.403	0.139	-2.894	0.00
Other	-0.741	0.185	-3.993	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.085	0.073	-1.170	0.24
Roman Catholic	-0.108	0.085	-1.273	0.20
Other religion	-0.534	0.112	-4.764	0.00
Life satisfaction (lifesat2)				
Above mode	-	-	-	-
Mode	0.013	0.068	0.197	0.84
Below mode	-0.146	0.071	-2.061	0.04
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.153	0.084	-1.831	0.07
A little	-0.327	0.082	-3.971	0.00
A lot	-0.573	0.098	-5.835	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.053	0.084	-0.636	0.52
Higher grade	-0.199	0.088	-2.270	0.02
Standard School Grade	-0.286	0.083	-3.455	0.00
Not qualifications	-0.495	0.117	-4.229	0.00
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.198	0.083	-2.393	0.02
3rd	-0.385	0.089	-4.332	0.00
2nd	-0.542	0.085	-6.388	0.00
Most deprived	-0.514	0.092	-5.579	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.341	0.087	-3.940	0.00
Very low activity	-0.449	0.092	-4.870	0.00
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.725	0.063	11.535	0.00
Light smokers	0.878	0.119	7.380	0.00
Moderate smokers	0.776	0.110	7.023	0.00
Heavy smokers	0.918	0.146	6.269	0.00

TABLE I.21 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE URBAN TRAINING DATASET K = 3 OR = 0.75 (smk3r75_dwu)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	0.039	0.126	0.308	0.76
Year				
2017	-	-	-	-
2018	-0.022	0.077	-0.289	0.77
2019	0.373	0.074	5.049	0.00
2021	0.480	0.075	6.364	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.875	0.050	-17.399	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.021	0.081	0.256	0.80
45-54	0.545	0.077	7.088	0.00
55-64	0.737	0.079	9.360	0.00
Birthplace (birthpla3)				
Scotland	-	-	-	-
Rest UK	0.211	0.132	1.600	0.11
Elsewhere	-0.140	0.166	-0.839	0.40
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.500	0.130	-3.850	0.00
Other	-0.823	0.173	-4.761	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.057	0.065	-0.870	0.38
Roman Catholic	-0.194	0.077	-2.501	0.01
Other religion	-0.623	0.102	-6.089	0.00
Life satisfaction (lifesat2)				
Above mode	-	-	-	-
Mode	0.107	0.062	1.730	0.08
Below mode	-0.094	0.067	-1.415	0.16
General health (genhelf2)				
Very good	-	-	-	-
Good	-0.019	0.061	-0.314	0.75
Fair	-0.149	0.090	-1.663	0.10
Bad_ Very bad	0.262	0.143	1.829	0.07
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.168	0.077	-2.179	0.03
A little	-0.383	0.080	-4.798	0.00
A lot	-0.663	0.110	-6.046	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.058	0.079	-0.733	0.46
Higher grade	-0.176	0.082	-2.135	0.03
Standard School Grade	-0.231	0.079	-2.915	0.00
Not qualifications	-0.326	0.110	-2.956	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.140	0.089	1.563	0.12
IIIN Skilled non-manual	-0.181	0.110	-1.645	0.10
IIIM Skilled manual	-0.061	0.108	-0.566	0.57
IV Semi-skilled manual	-0.039	0.117	-0.328	0.74
Unskilled or other	-0.247	0.150	-1.641	0.10
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.303	0.076	-3.999	0.00
3rd	-0.417	0.081	-5.146	0.00
2nd	-0.523	0.077	-6.750	0.00
Most deprived	-0.483	0.085	-5.673	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.419	0.079	-5.300	0.00
Very low activity	-0.531	0.086	-6.149	0.00
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.784	0.057	13.750	0.00
Light smokers	0.879	0.109	8.052	0.00
Moderate smokers	0.827	0.100	8.282	0.00
Heavy smokers	0.983	0.134	7.345	0.00

TABLE I.22 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE URBAN TRAINING DATASET K = 3 OR = 1 (smk3r1_dwu)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	0.324	0.117	2.761	0.01
Year				
2017	-	-	-	-
2018	0.020	0.072	0.278	0.78
2019	0.458	0.069	6.606	0.00
2021	0.521	0.071	7.360	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.873	0.047	-18.601	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.170	0.075	2.278	0.02
45-54	0.609	0.072	8.428	0.00
55-64	0.847	0.074	11.493	0.00
Birthplace (birthpla3)				
Scotland	-	-	-	-
Rest UK	0.240	0.129	1.864	0.06
Elsewhere	-0.269	0.156	-1.725	0.08
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.557	0.126	-4.411	0.00
Other	-0.777	0.161	-4.824	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.159	0.062	-2.581	0.01
Roman Catholic	-0.215	0.072	-2.996	0.00
Other religion	-0.709	0.095	-7.435	0.00
Life satisfaction (lifesat2)				
Above mode	-	-	-	-
Mode	0.057	0.058	0.989	0.32
Below mode	-0.098	0.062	-1.568	0.12
General health (genhelf2)				
Very good	-	-	-	-
Good	0.018	0.057	0.319	0.75
Fair	-0.145	0.085	-1.706	0.09
Bad_ Very bad	0.167	0.135	1.232	0.22
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.290	0.074	-3.946	0.00
A little	-0.423	0.075	-5.651	0.00
A lot	-0.607	0.102	-5.978	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.103	0.074	-1.391	0.16
Higher grade	-0.177	0.077	-2.310	0.02
Standard School Grade	-0.247	0.074	-3.346	0.00
Not qualifications	-0.442	0.104	-4.244	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.110	0.083	1.316	0.19
IIIN Skilled non-manual	-0.241	0.102	-2.365	0.02
IIIM Skilled manual	-0.076	0.101	-0.751	0.45
IV Semi-skilled manual	-0.020	0.110	-0.183	0.85
Unskilled or other	-0.250	0.140	-1.790	0.07
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.247	0.071	-3.499	0.00
3rd	-0.438	0.076	-5.796	0.00
2nd	-0.523	0.072	-7.222	0.00
Most deprived	-0.413	0.079	-5.224	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.491	0.074	-6.654	0.00
Very low activity	-0.569	0.081	-6.987	0.00
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.727	0.053	13.713	0.00
Light smokers	0.758	0.103	7.338	0.00
Moderate smokers	0.677	0.095	7.125	0.00
Heavy smokers	0.824	0.128	6.429	0.00

TABLE I.23 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE URBAN TRAINING DATASET K = 5 OR = 0.5 (smk5r50_dwu)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.256	0.138	-1.848	0.06
Year				
2017	-	-	-	-
2018	0.009	0.084	0.104	0.92
2019	0.224	0.081	2.754	0.01
2021	0.258	0.084	3.084	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.870	0.056	-15.602	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.105	0.090	1.168	0.24
45-54	0.527	0.086	6.162	0.00
55-64	0.718	0.087	8.257	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.254	0.088	-2.898	0.00
Other	-0.911	0.118	-7.732	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.060	0.073	-0.825	0.41
Roman Catholic	-0.089	0.085	-1.054	0.29
Other religion	-0.519	0.113	-4.585	0.00
Life satisfaction (lifesat2)				
Above mode	-	-	-	-
Mode	-0.004	0.068	-0.061	0.95
Below mode	-0.163	0.071	-2.284	0.02
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.165	0.084	-1.955	0.05
A little	-0.330	0.083	-3.984	0.00
A lot	-0.501	0.098	-5.116	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.057	0.087	-0.652	0.51
Higher grade	-0.143	0.090	-1.584	0.11
Standard School Grade	-0.260	0.088	-2.954	0.00
Not qualifications	-0.368	0.120	-3.058	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.071	0.098	0.725	0.47
IIIN Skilled non-manual	-0.215	0.122	-1.765	0.08
IIIM Skilled manual	-0.069	0.119	-0.581	0.56
IV Semi-skilled manual	-0.087	0.130	-0.673	0.50
Unskilled or other	-0.109	0.161	-0.674	0.50
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.258	0.084	-3.059	0.00
3rd	-0.333	0.090	-3.707	0.00
2nd	-0.426	0.085	-5.020	0.00
Most deprived	-0.434	0.094	-4.635	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.371	0.087	-4.236	0.00
Very low activity	-0.492	0.093	-5.317	0.00
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.712	0.063	11.297	0.00
Light smokers	0.911	0.118	7.740	0.00
Moderate smokers	0.679	0.113	6.010	0.00
Heavy smokers	0.916	0.146	6.260	0.00

TABLE I.24 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE URBAN TRAINING DATASET K = 5 OR = 0.75 (smk5r0.75_dwu)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.014	0.128	-0.107	0.91
Year				
2017	-	-	-	-
2018	0.011	0.077	0.140	0.89
2019	0.379	0.074	5.084	0.00
2021	0.473	0.076	6.222	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.902	0.050	-17.888	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.195	0.082	2.393	0.02
45-54	0.689	0.078	8.849	0.00
55-64	0.845	0.080	10.604	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.335	0.081	-4.154	0.00
Other	-1.010	0.108	-9.348	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.060	0.065	-0.912	0.36
Roman Catholic	-0.134	0.077	-1.753	0.08
Other religion	-0.761	0.107	-7.102	0.00
Life satisfaction (lifesat2)				
Above mode	-	-	-	-
Mode	0.078	0.062	1.260	0.21
Below mode	-0.128	0.067	-1.924	0.05
General health (genhelf2)				
Very good	-	-	-	-
Good	-0.112	0.061	-1.847	0.06
Fair	-0.225	0.091	-2.484	0.01
Bad_ Very bad	0.180	0.144	1.252	0.21
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.195	0.078	-2.504	0.01
A little	-0.351	0.080	-4.369	0.00
A lot	-0.582	0.110	-5.314	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.120	0.079	-1.520	0.13
Higher grade	-0.193	0.082	-2.346	0.02
Standard School Grade	-0.303	0.080	-3.793	0.00
Not qualifications	-0.439	0.112	-3.931	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.156	0.090	1.736	0.08
IIIN Skilled non-manual	-0.073	0.110	-0.661	0.51
IIIM Skilled manual	-0.036	0.110	-0.332	0.74
IV Semi-skilled manual	0.069	0.118	0.589	0.56
Unskilled or other	-0.155	0.151	-1.026	0.30
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.307	0.076	-4.024	0.00
3rd	-0.352	0.081	-4.362	0.00
2nd	-0.486	0.077	-6.276	0.00
Most deprived	-0.447	0.085	-5.239	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.410	0.079	-5.193	0.00
Very low activity	-0.544	0.087	-6.238	0.00
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.756	0.057	13.255	0.00
Light smokers	0.897	0.110	8.152	0.00
Moderate smokers	0.743	0.102	7.297	0.00
Heavy smokers	0.869	0.136	6.368	0.00

TABLE I.25 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE URBAN TRAINING DATASET K = 5 OR = 1 (smk5r100_dwu)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	0.215	0.116	1.850	0.06
Year				
2017	-	-	-	-
2018	0.095	0.073	1.295	0.20
2019	0.505	0.071	7.158	0.00
2021	0.644	0.072	8.964	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.850	0.047	-18.094	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.155	0.075	2.063	0.04
45-54	0.607	0.072	8.397	0.00
55-64	0.818	0.074	11.041	0.00
Birthplace (birthpla3)				
Scotland	-	-	-	-
Rest UK	0.154	0.129	1.196	0.23
Elsewhere	-0.256	0.161	-1.597	0.11
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.509	0.126	-4.039	0.00
Other	-0.901	0.167	-5.404	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.103	0.061	-1.682	0.09
Roman Catholic	-0.196	0.072	-2.709	0.01
Other religion	-0.858	0.100	-8.598	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	-0.053	0.056	-0.948	0.34
Fair	-0.227	0.083	-2.720	0.01
Bad_ Very bad	0.082	0.133	0.617	0.54
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.202	0.073	-2.755	0.01
A little	-0.384	0.075	-5.122	0.00
A lot	-0.640	0.103	-6.235	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.154	0.075	-2.061	0.04
Higher grade	-0.145	0.077	-1.879	0.06
Standard School Grade	-0.285	0.075	-3.823	0.00
Not qualifications	-0.324	0.103	-3.140	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.143	0.084	1.704	0.09
IIIN Skilled non-manual	-0.206	0.103	-1.994	0.05
IIIM Skilled manual	0.023	0.102	0.228	0.82
IV Semi-skilled manual	0.015	0.111	0.136	0.89
Unskilled or other	-0.125	0.138	-0.900	0.37
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.234	0.071	-3.295	0.00
3rd	-0.361	0.076	-4.765	0.00
2nd	-0.497	0.073	-6.811	0.00
Most deprived	-0.444	0.080	-5.571	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.538	0.075	-7.169	0.00
Very low activity	-0.549	0.082	-6.715	0.00
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.722	0.053	13.569	0.00
Light smokers	0.806	0.104	7.785	0.00
Moderate smokers	0.666	0.096	6.932	0.00
Heavy smokers	0.896	0.126	7.109	0.00

TABLE I.26 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE URBAN TRAINING DATASET K = 10 OR = 0.5 (smk10r50_dwu)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.234	0.141	-1.656	0.10
Year				
2017	-	-	-	-
2018	-0.062	0.084	-0.732	0.46
2019	0.207	0.081	2.543	0.01
2021	0.294	0.083	3.537	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.880	0.056	-15.675	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.071	0.093	0.762	0.45
45-54	0.519	0.090	5.796	0.00
55-64	0.702	0.093	7.512	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.340	0.089	-3.816	0.00
Other	-0.947	0.119	-7.980	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.078	0.073	-1.069	0.29
Roman Catholic	-0.149	0.086	-1.731	0.08
Other religion	-0.487	0.113	-4.322	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	0.104	0.084	1.248	0.21
Single	-0.142	0.083	-1.713	0.09
Separated & Widowed	-0.005	0.092	-0.054	0.96
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.139	0.085	-1.639	0.10
A little	-0.251	0.081	-3.095	0.00
A lot	-0.599	0.098	-6.132	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.119	0.088	-1.352	0.18
Higher grade	-0.176	0.091	-1.933	0.05
Standard School Grade	-0.270	0.088	-3.071	0.00
Not qualifications	-0.356	0.121	-2.947	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.058	0.098	0.594	0.55
IIIN Skilled non-manual	-0.216	0.122	-1.772	0.08
IIIM Skilled manual	-0.134	0.120	-1.115	0.27
IV Semi-skilled manual	-0.148	0.131	-1.129	0.26
Unskilled or other	-0.148	0.163	-0.905	0.37
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.299	0.084	-3.567	0.00
3rd	-0.361	0.090	-4.026	0.00
2nd	-0.541	0.086	-6.286	0.00
Most deprived	-0.498	0.095	-5.257	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.225	0.085	-2.649	0.01
Very low activity	-0.415	0.093	-4.472	0.00
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.764	0.063	12.056	0.00
Light smokers	0.892	0.121	7.377	0.00
Moderate smokers	0.856	0.111	7.694	0.00
Heavy smokers	1.004	0.146	6.895	0.00

TABLE I.27 COEFICIENTS FROM CROSS-VALIDATION IN THE SMOTE URBAN TRAINING DATASET K = 10 OR = 0.75 (smk10r75_dwu)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	0.048	0.127	0.375	0.71
Year				
2017	-	-	-	-
2018	0.003	0.076	0.035	0.97
2019	0.274	0.074	3.690	0.00
2021	0.467	0.075	6.221	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.894	0.050	-17.735	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.176	0.081	2.164	0.03
45-54	0.663	0.078	8.541	0.00
55-64	0.817	0.080	10.251	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.360	0.081	-4.467	0.00
Other	-0.984	0.106	-9.313	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.059	0.065	-0.899	0.37
Roman Catholic	-0.145	0.077	-1.875	0.06
Other religion	-0.558	0.102	-5.486	0.00
Life satisfaction (lifesat2)				
Above mode	-	-	-	-
Mode	0.065	0.062	1.046	0.30
Below mode	-0.101	0.067	-1.523	0.13
General health (genhelf2)				
Very good	-	-	-	-
Good	-0.026	0.060	-0.430	0.67
Fair	-0.240	0.091	-2.623	0.01
Bad_ Very bad	0.152	0.145	1.048	0.29
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.204	0.077	-2.644	0.01
A little	-0.395	0.080	-4.921	0.00
A lot	-0.629	0.111	-5.679	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.061	0.079	-0.777	0.44
Higher grade	-0.208	0.083	-2.492	0.01
Standard School Grade	-0.300	0.080	-3.730	0.00
Not qualifications	-0.421	0.112	-3.770	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.136	0.089	1.537	0.12
IIIN Skilled non-manual	-0.201	0.110	-1.821	0.07
IIIM Skilled manual	-0.008	0.109	-0.075	0.94
IV Semi-skilled manual	-0.001	0.118	-0.007	0.99
Unskilled or other	-0.051	0.148	-0.343	0.73
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.369	0.077	-4.822	0.00
3rd	-0.356	0.079	-4.485	0.00
2nd	-0.552	0.077	-7.138	0.00
Most deprived	-0.539	0.085	-6.310	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.411	0.079	-5.211	0.00
Very low activity	-0.441	0.086	-5.148	0.00
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.755	0.057	13.229	0.00
Light smokers	0.875	0.110	7.980	0.00
Moderate smokers	0.684	0.103	6.646	0.00
Heavy smokers	1.055	0.132	7.982	0.00

TABLE I.28 COEFICIENTS FROM CROSS-VALIDATION IN THE SMOTE URBAN TRAINING DATASET K = 10 OR = 1 (smk10r100_dwu)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	0.199	0.123	1.620	0.11
Year				
2017	-	-	-	-
2018	0.052	0.073	0.712	0.48
2019	0.410	0.070	5.813	0.00
2021	0.628	0.071	8.820	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.858	0.047	-18.094	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.269	0.078	3.442	0.00
45-54	0.710	0.077	9.236	0.00
55-64	0.912	0.081	11.305	0.00
Birthplace (birthpla3)				
Scotland	-	-	-	-
Rest UK	0.149	0.124	1.199	0.23
Elsewhere	-0.198	0.160	-1.240	0.21
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	-0.574	0.122	-4.713	0.00
Other	-1.008	0.167	-6.051	0.00
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.118	0.062	-1.914	0.06
Roman Catholic	-0.269	0.074	-3.635	0.00
Other religion	-0.741	0.098	-7.530	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	0.078	0.072	1.085	0.28
Single	-0.106	0.070	-1.521	0.13
Separated & Widowed	-0.173	0.080	-2.158	0.03
General health (genhelf2)				
Very good	-	-	-	-
Good	-0.095	0.056	-1.706	0.09
Fair	-0.308	0.084	-3.665	0.00
Bad_ Very bad	0.085	0.133	0.644	0.52
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.269	0.074	-3.642	0.00
A little	-0.417	0.075	-5.548	0.00
A lot	-0.529	0.101	-5.226	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.084	0.074	-1.129	0.26
Higher grade	-0.263	0.078	-3.366	0.00
Standard School Grade	-0.230	0.075	-3.086	0.00
Not qualifications	-0.362	0.105	-3.453	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.257	0.084	3.041	0.00
IIIN Skilled non-manual	-0.149	0.104	-1.430	0.15
IIIM Skilled manual	0.064	0.103	0.626	0.53
IV Semi-skilled manual	0.036	0.112	0.319	0.75
Unskilled or other	-0.032	0.141	-0.230	0.82
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.292	0.071	-4.125	0.00
3rd	-0.425	0.076	-5.605	0.00
2nd	-0.533	0.072	-7.352	0.00
Most deprived	-0.536	0.082	-6.561	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.449	0.074	-6.052	0.00
Very low activity	-0.552	0.082	-6.725	0.00

Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.723	0.054	13.500	0.00
Light smokers	0.847	0.104	8.123	0.00
Moderate smokers	0.655	0.097	6.719	0.00
Heavy smokers	1.011	0.125	8.072	0.00

TABLE I.29 COEFICIENTS FROM CROSS-VALIDATION IN THE RURAL TRAINING DATASET (dwr_train)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-1.026	0.236	-4.346	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.726	0.129	-5.648	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.343	0.239	1.435	0.15
45-54	0.538	0.228	2.365	0.02
55-64	0.566	0.229	2.475	0.01
Birthplace				
Scotland	-	-	-	-
Rest UK	-0.477	0.276	-1.727	0.08
Elsewhere	-1.294	0.524	-2.467	0.01
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	0.545	0.279	1.954	0.05
Other	-0.080	0.504	-0.158	0.87
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.293	0.171	-1.715	0.09
Roman Catholic	0.123	0.235	0.523	0.60
Other religion	-0.525	0.237	-2.217	0.03
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.129	0.187	-0.688	0.49
A little	-0.390	0.188	-2.075	0.04
A lot	-1.069	0.242	-4.420	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.113	0.191	-0.591	0.55
Higher grade	-0.316	0.191	-1.654	0.10
Standard School Grade	-0.413	0.194	-2.128	0.03
Not qualifications	-0.798	0.270	-2.952	0.00
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.564	0.143	3.943	0.00
Light smokers	1.001	0.287	3.490	0.00
Moderate smokers	0.902	0.259	3.482	0.00

TABLE I.30 COEFICIENTS FROM CROSS-VALIDATION IN THE OVERSAMPLING RURAL TRAINING DATASET 30% (up30_dwr)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.456	0.311	-1.465	0.14
Sex (sex)				
Male	-	-	-	-
Female	-0.781	0.118	-6.611	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.381	0.229	1.662	0.10
45-54	0.649	0.214	3.028	0.00
55-64	0.609	0.215	2.835	0.00
Birthplace (birthpla3)				
Scotland	-	-	-	-
Rest UK	-0.527	0.251	-2.096	0.04
Elsewhere	-1.018	0.451	-2.255	0.02
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	0.626	0.254	2.460	0.01
Other	-0.723	0.503	-1.438	0.15
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.274	0.172	-1.589	0.11
A little	-0.571	0.179	-3.198	0.00
A lot	-1.081	0.231	-4.679	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	-0.165	0.204	-0.806	0.42
IIIN Skilled non-manual	-0.710	0.274	-2.592	0.01
IIIM Skilled manual	-0.475	0.240	-1.981	0.05
IV Semi-skilled manual	-0.593	0.262	-2.264	0.02
V Unskilled manual & other	-0.801	0.351	-2.280	0.02
Employment Status (neconacb)				
In employment	-	-	-	-
ILO unemployed & Inactive	-0.291	0.168	-1.729	0.08
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.178	0.198	-0.894	0.37
3rd	-0.373	0.205	-1.825	0.07
2nd	-0.709	0.279	-2.546	0.01
Most deprived	-0.872	0.491	-1.777	0.08
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.769	0.131	5.882	0.00
Light smokers	1.219	0.274	4.455	0.00
Moderate smokers	1.116	0.236	4.730	0.00
Heavy smokers	1.331	0.311	4.279	0.00

TABLE I.31 COEFICIENTS FROM CROSS-VALIDATION IN THE OVERSAMPLING RURAL TRAINING DATASET 40% (up40_dwr)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.510	0.204	-2.501	0.01
Sex (sex)				
Male	-	-	-	-
Female	-0.884	0.103	-8.590	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.545	0.191	2.857	0.00
45-54	0.591	0.185	3.188	0.00
55-64	0.695	0.189	3.681	0.00
Birthplace (birthpla3)				
Scotland	-	-	-	-
Rest UK	-0.564	0.228	-2.475	0.01
Elsewhere	-0.843	0.361	-2.339	0.02
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	0.765	0.229	3.345	0.00
Other	-0.353	0.381	-0.928	0.35
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.130	0.134	-0.973	0.33
Roman Catholic	0.205	0.188	1.092	0.27
Other religion	-0.438	0.178	-2.458	0.01
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	-0.024	0.153	-0.157	0.87
Single	0.053	0.162	0.327	0.74
Separated & Widowed	-0.548	0.189	-2.899	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.316	0.119	2.649	0.01
Fair	0.290	0.173	1.679	0.09
Bad_ Very bad	0.232	0.309	0.751	0.45
LTC (limitac_h)				
`	-	-	-	-
Not at all	-0.285	0.155	-1.838	0.07
A little	-0.437	0.158	-2.766	0.01
A lot	-0.963	0.227	-4.248	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	0.011	0.151	0.073	0.94
Higher grade	-0.364	0.152	-2.393	0.02
Standard School Grade	-0.494	0.154	-3.198	0.00
Not qualifications	-0.903	0.213	-4.244	0.00
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.563	0.114	4.944	0.00
Light smokers	0.859	0.246	3.498	0.00
Moderate smokers	1.015	0.209	4.857	0.00
Heavy smokers	1.371	0.285	4.804	0.00

TABLE I.32 COEFICIENTS FROM CROSS-VALIDATION IN THE OVERSAMPLING RURAL TRAINING DATASET 50% (up50_dwr)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	0.211	0.226	0.931	0.35
Sex (sex)				
Male	-	-	-	-
Female	-0.784	0.093	-8.393	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.499	0.171	2.927	0.00
45-54	0.660	0.167	3.949	0.00
55-64	0.812	0.169	4.794	0.00
Birthplace (birthpla3)				
Scotland	-	-	-	-
Rest UK	-0.622	0.200	-3.106	0.00
Elsewhere	-1.870	0.387	-4.830	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	0.585	0.205	2.858	0.00
Other	0.180	0.355	0.508	0.61
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.338	0.123	-2.751	0.01
Roman Catholic	0.196	0.175	1.119	0.26
Other religion	-0.518	0.164	-3.167	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	0.172	0.135	1.276	0.20
Single	0.226	0.151	1.496	0.13
Separated & Widowed	-0.563	0.173	-3.263	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.465	0.107	4.327	0.00
Fair	0.321	0.155	2.070	0.04
Bad_ Very bad	0.603	0.287	2.103	0.04
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.131	0.136	-0.968	0.33
A little	-0.684	0.145	-4.719	0.00
A lot	-1.222	0.214	-5.708	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.261	0.142	-1.841	0.07
Higher grade	-0.246	0.134	-1.835	0.07
Standard School Grade	-0.351	0.138	-2.543	0.01
Not qualifications	-1.055	0.206	-5.129	0.00
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.302	0.158	-1.912	0.06
3rd	-0.468	0.162	-2.895	0.00
2nd	-0.599	0.211	-2.843	0.00
Most deprived	-1.087	0.391	-2.778	0.01

Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.498	0.104	4.802	0.00
Light smokers	0.912	0.223	4.082	0.00
Moderate smokers	1.047	0.194	5.410	0.00
Heavy smokers	1.209	0.268	4.504	0.00

TABLE I.33 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE TRAINING RURAL DATASET k = 3 / or = 0.5 (smk3r50_dwr)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.506	0.224	-2.252	0.02
Sex (sex)				
Male	-	-	-	-
Female	-0.801	0.113	-7.077	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.287	0.216	1.327	0.18
45-54	0.706	0.207	3.417	0.00
55-64	0.666	0.212	3.150	0.00
Birthplace (birthpla3)				
Scotland	-	-	-	-
Rest UK	-0.593	0.257	-2.301	0.02
Elsewhere	-1.288	0.463	-2.784	0.01
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	0.522	0.261	2.001	0.05
Other	-0.283	0.464	-0.609	0.54
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.426	0.150	-2.830	0.00
Roman Catholic	-0.171	0.223	-0.768	0.44
Other religion	-0.710	0.214	-3.325	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	-0.217	0.172	-1.262	0.21
Single	-0.199	0.191	-1.039	0.30
Separated & Widowed	-0.755	0.223	-3.390	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.447	0.129	3.472	0.00
Fair	0.144	0.198	0.726	0.47
Bad_ Very bad	0.362	0.357	1.016	0.31
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.305	0.170	-1.795	0.07
A little	-0.530	0.176	-3.002	0.00
A lot	-1.218	0.274	-4.445	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.294	0.172	-1.710	0.09
Higher grade	-0.455	0.167	-2.720	0.01
Standard School Grade	-0.552	0.170	-3.253	0.00
Not qualifications	-1.016	0.245	-4.143	0.00
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.604	0.124	4.857	0.00
Light smokers	0.898	0.278	3.231	0.00
Moderate smokers	0.900	0.243	3.698	0.00
Heavy smokers	1.278	0.319	4.001	0.00

TABLE I.34 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE TRAINING RURAL DATASET k = 3 / or = 0.75 (smk3r75_dwr)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.203	0.223	-0.911	0.36
Survey year (syear)				
2017 -	-	-	-	
2018	0.292	0.156	1.868	0.06
2019	0.335	0.154	2.176	0.03
2021	0.503	0.154	3.265	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.730	0.102	-7.141	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.234	0.188	1.245	0.21
45-54	0.533	0.180	2.968	0.00
55-64	0.622	0.187	3.324	0.00
Birthplace (birthpla3)				
Scotland	-	-	-	-
Rest UK	-0.711	0.232	-3.064	0.00
Elsewhere	-1.948	0.441	-4.412	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	0.676	0.236	2.859	0.00
Other	0.097	0.408	0.239	0.81
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.381	0.134	-2.849	0.00
Roman Catholic	-0.226	0.202	-1.116	0.26
Other religion	-0.809	0.192	-4.217	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	-0.247	0.152	-1.622	0.10
Single	-0.206	0.173	-1.190	0.23
Separated & Widowed	-0.669	0.191	-3.495	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.521	0.116	4.489	0.00
Fair	0.193	0.178	1.082	0.28
Bad_ Very bad	0.663	0.339	1.954	0.05
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.473	0.155	-3.042	0.00
A little	-0.695	0.161	-4.327	0.00
A lot	-1.337	0.259	-5.171	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.334	0.153	-2.182	0.03
Higher grade	-0.442	0.148	-2.981	0.00
Standard School Grade	-0.763	0.159	-4.794	0.00
Not qualifications	-1.186	0.235	-5.052	0.00
Employment Status (neconacb)				
In employment	-	-	-	-
ILO unemployed & Inactive	-0.296	0.151	-1.962	0.05
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.444	0.113	3.938	0.00
Light smokers	0.790	0.257	3.075	0.00
Moderate smokers	0.957	0.218	4.389	0.00
Heavy smokers	1.216	0.302	4.022	0.00

TABLE I.35 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE TRAINING RURAL DATASET k = 3 / or = 1 (smk3r100_dwr)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	0.224	0.259	0.867	0.39
Survey year (syear)				
2017	-	-	-	-
2018	0.323	0.150	2.155	0.03
2019	0.426	0.147	2.893	0.00
2021	0.578	0.150	3.851	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.713	0.097	-7.350	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.264	0.176	1.497	0.13
45-54	0.483	0.170	2.837	0.00
55-64	0.653	0.181	3.617	0.00
Birthplace (birthpla3)				
Scotland	-	-	-	-
Rest UK	-0.795	0.224	-3.542	0.00
Elsewhere	-1.700	0.407	-4.182	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	0.621	0.228	2.722	0.01
Other	-0.792	0.433	-1.830	0.07
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.554	0.128	-4.329	0.00
Roman Catholic	-0.100	0.190	-0.525	0.60
Other religion	-1.194	0.197	-6.048	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	-0.294	0.144	-2.048	0.04
Single	-0.189	0.166	-1.136	0.26
Separated & Widowed	-0.907	0.191	-4.735	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.424	0.112	3.793	0.00
Fair	0.079	0.173	0.456	0.65
Bad_ Very bad	0.974	0.318	3.059	0.00
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.556	0.150	-3.696	0.00
A little	-0.675	0.152	-4.449	0.00
A lot	-1.200	0.241	-4.982	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.291	0.145	-2.004	0.05
Higher grade	-0.432	0.143	-3.025	0.00
Standard School Grade	-0.544	0.148	-3.672	0.00
Not qualifications	-0.930	0.223	-4.175	0.00
Employment Status (neconacb)				
In employment	-	-	-	-
ILO unemployed & Inactive	-0.217	0.141	-1.536	0.12
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	-0.063	0.170	-0.372	0.71
3rd	-0.149	0.172	-0.870	0.38
2nd	-0.491	0.228	-2.155	0.03
Most deprived	-1.715	0.532	-3.223	0.00
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.137	0.151	-0.905	0.37
Very low activity	-0.448	0.181	-2.469	0.01
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.627	0.106	5.936	0.00
Light smokers	0.918	0.244	3.765	0.00
Moderate smokers	1.039	0.214	4.846	0.00
Heavy smokers	1.155	0.300	3.845	0.00

TABLE I.36 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE TRAINING RURAL DATASET k = 5 / or = 0.5 (smk5r50_dwr)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.976	0.306	-3.191	0.00
Survey year (syear)				
2017	-	-	-	-
2018	0.416	0.176	2.358	0.02
2019	0.432	0.172	2.504	0.01
2021	0.460	0.177	2.604	0.01
Sex (sex)				
Male	-	-	-	-
Female	-0.693	0.113	-6.122	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.405	0.218	1.863	0.06
45-54	0.697	0.210	3.324	0.00
55-64	0.777	0.214	3.624	0.00
Birthplace (birthpla3)				
Scotland	-	-	-	-
Rest UK	-0.477	0.247	-1.930	0.05
Elsewhere	-1.367	0.473	-2.888	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	0.419	0.253	1.659	0.10
Other	-0.313	0.475	-0.659	0.51
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.398	0.151	-2.626	0.01
Roman Catholic	-0.001	0.218	-0.006	1.00
Other religion	-0.696	0.214	-3.246	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	-0.066	0.168	-0.395	0.69
Single	-0.022	0.191	-0.114	0.91
Separated & Widowed	-0.758	0.227	-3.343	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.276	0.129	2.141	0.03
Fair	-0.038	0.200	-0.189	0.85
Bad_ Very bad	-0.119	0.371	-0.322	0.75
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.345	0.173	-1.999	0.05
A little	-0.442	0.175	-2.519	0.01
A lot	-1.017	0.270	-3.770	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.166	0.170	-0.976	0.33
Higher grade	-0.421	0.178	-2.366	0.02
Standard School Grade	-0.469	0.181	-2.584	0.01
Not qualifications	-0.897	0.264	-3.399	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.125	0.203	0.615	0.54
IIIN Skilled non-manual	-0.375	0.270	-1.390	0.16
IIIM Skilled manual	0.072	0.242	0.299	0.76
IV Semi-skilled manual	-0.348	0.276	-1.263	0.21
V Unskilled manual & other	-0.207	0.343	-0.602	0.55
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.602	0.124	4.838	0.00
Light smokers	0.934	0.273	3.420	0.00
Moderate smokers	0.944	0.250	3.783	0.00
Heavy smokers	1.292	0.331	3.903	0.00

TABLE I.37 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE TRAINING RURAL DATASET k = 5 / or = 0.75 (smk5r75_dwr)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.498	0.313	-1.590	0.11
Survey year (syear)				
2017	-	-	-	-
2018	0.269	0.162	1.659	0.10
2019	0.361	0.157	2.298	0.02
2021	0.630	0.157	4.000	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.701	0.104	-6.744	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.378	0.198	1.905	0.06
45-54	0.634	0.192	3.309	0.00
55-64	0.652	0.199	3.278	0.00
Birthplace (birthpla3)				
Scotland	-	-	-	-
Rest UK	-0.733	0.245	-2.991	0.00
Elsewhere	-1.317	0.416	-3.166	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	0.620	0.248	2.502	0.01
Other	-0.476	0.436	-1.092	0.27
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.623	0.140	-4.442	0.00
Roman Catholic	-0.066	0.201	-0.328	0.74
Other religion	-0.933	0.199	-4.696	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	-0.425	0.160	-2.651	0.01
Single	-0.108	0.179	-0.607	0.54
Separated & Widowed	-0.964	0.212	-4.553	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.430	0.119	3.620	0.00
Fair	0.011	0.189	0.060	0.95
Bad_ Very bad	0.274	0.360	0.760	0.45
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.383	0.158	-2.423	0.02
A little	-0.465	0.162	-2.865	0.00
A lot	-1.262	0.267	-4.718	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.157	0.158	-0.995	0.32
Higher grade	-0.530	0.165	-3.207	0.00
Standard School Grade	-0.457	0.166	-2.762	0.01
Not qualifications	-0.682	0.245	-2.783	0.01
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.121	0.183	0.660	0.51
IIIN Skilled non-manual	-0.475	0.249	-1.905	0.06
IIIM Skilled manual	0.013	0.223	0.058	0.95
IV Semi-skilled manual	-0.243	0.250	-0.971	0.33
V Unskilled manual & other	-0.090	0.308	-0.292	0.77
Employment Status (neconacb)				
In employment	-	-	-	-
ILO unemployed & Inactive	-0.251	0.154	-1.635	0.10
SIMD quintile (SIMD20_RPa)				
Least deprived				
4th	0.163	0.187	0.873	0.38
3rd	0.209	0.191	1.095	0.27
2nd	-0.216	0.253	-0.852	0.39
Most deprived	-0.653	0.484	-1.347	0.18
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.565	0.114	4.961	0.00
Light smokers	1.159	0.251	4.623	0.00
Moderate smokers	0.744	0.244	3.046	0.00
Heavy smokers	1.513	0.309	4.891	0.00

TABLE I.38 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE TRAINING RURAL DATASET k = 5 / or = 1 (smk5r100_dwr)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.483	0.302	-1.599	0.11
Survey year (syear)				
2017	-	-	-	-
2018	0.305	0.152	2.010	0.04
2019	0.385	0.149	2.594	0.01
2021	0.701	0.148	4.735	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.749	0.099	-7.601	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.533	0.188	2.832	0.00
45-54	0.854	0.182	4.689	0.00
55-64	0.815	0.188	4.331	0.00
Birthplace (birthpla3)				
Scotland	-	-	-	-
Rest UK	-0.844	0.230	-3.671	0.00
Elsewhere	-1.932	0.423	-4.563	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	0.594	0.234	2.539	0.01
Other	-0.108	0.403	-0.267	0.79
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.608	0.130	-4.674	0.00
Roman Catholic	-0.494	0.212	-2.332	0.02
Other religion	-0.766	0.177	-4.328	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	-0.352	0.149	-2.358	0.02
Single	-0.138	0.172	-0.802	0.42
Separated & Widowed	-0.943	0.197	-4.782	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.490	0.111	4.422	0.00
Fair	0.135	0.176	0.770	0.44
Bad_ Very bad	0.617	0.351	1.757	0.08
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.429	0.149	-2.890	0.00
A little	-0.633	0.150	-4.215	0.00
A lot	-1.465	0.259	-5.652	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.216	0.149	-1.451	0.15
Higher grade	-0.448	0.153	-2.921	0.00
Standard School Grade	-0.542	0.160	-3.398	0.00
Not qualifications	-0.849	0.237	-3.586	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.358	0.179	2.000	0.05
IIIN Skilled non-manual	-0.387	0.241	-1.608	0.11
IIIM Skilled manual	0.045	0.218	0.208	0.84
IV Semi-skilled manual	-0.078	0.241	-0.322	0.75
V Unskilled manual & other	-0.106	0.305	-0.346	0.73
SIMD quintile (SIMD20_RPa)				
Least deprived	-	-	-	-
4th	0.172	0.174	0.991	0.32
3rd	0.226	0.177	1.272	0.20
2nd	-0.193	0.234	-0.827	0.41
Most deprived	-1.197	0.538	-2.225	0.03
Physical activity level (adt10gptw)				
Meet recommendations	-	-	-	-
Low activity	-0.082	0.152	-0.538	0.59
Very low activity	-0.417	0.191	-2.181	0.03

Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.469	0.108	4.339	0.00
Light smokers	1.071	0.250	4.277	0.00
Moderate smokers	1.201	0.217	5.530	0.00
Heavy smokers	1.422	0.305	4.667	0.00

TABLE I.39 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE TRAINING RURAL DATASET k = 10 / or = 0.50 (smk10r50_dwr)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.426	0.220	-1.934	0.05
Sex (sex)				
Male	-	-	-	-
Female	-0.816	0.113	-7.221	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.407	0.215	1.893	0.06
45-54	0.728	0.206	3.525	0.00
55-64	0.758	0.211	3.594	0.00
Birthplace (birthpla3)				
Scotland	-	-	-	-
Rest UK	-0.530	0.251	-2.106	0.04
Elsewhere	-1.518	0.473	-3.209	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	0.508	0.255	1.990	0.05
Other	-0.106	0.448	-0.237	0.81
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.533	0.152	-3.501	0.00
Roman Catholic	-0.088	0.216	-0.408	0.68
Other religion	-0.803	0.215	-3.735	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	-0.202	0.171	-1.182	0.24
Single	-0.080	0.187	-0.428	0.67
Separated & Widowed	-0.715	0.223	-3.200	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.368	0.128	2.883	0.00
Fair	0.002	0.201	0.010	0.99
Bad_ Very bad	0.107	0.375	0.285	0.78
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.306	0.171	-1.791	0.07
A little	-0.475	0.175	-2.714	0.01
A lot	-1.146	0.280	-4.100	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.422	0.174	-2.422	0.02
Higher grade	-0.458	0.166	-2.768	0.01
Standard School Grade	-0.616	0.172	-3.585	0.00
Not qualifications	-1.039	0.247	-4.205	0.00
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.490	0.125	3.924	0.00
Light smokers	0.962	0.270	3.557	0.00
Moderate smokers	0.805	0.244	3.303	0.00
Heavy smokers	0.928	0.334	2.777	0.01

TABLE I.40 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE TRAINING RURAL DATASET k = 10 / or = 0.75 (smk10r75_dwr)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.509	0.282	-1.807	0.07
Survey year (syear)				
2017	-	-	-	-
2018	0.401	0.160	2.507	0.01
2019	0.345	0.158	2.183	0.03
2021	0.512	0.160	3.201	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.792	0.104	-7.597	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.432	0.198	2.182	0.03
45-54	0.741	0.191	3.882	0.00
55-64	0.897	0.197	4.550	0.00
Birthplace (birthpla3)				
Scotland	-	-	-	-
Rest UK	-0.710	0.243	-2.917	0.00
Elsewhere	-1.511	0.448	-3.373	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	0.484	0.248	1.950	0.05
Other	-0.618	0.464	-1.330	0.18
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.586	0.139	-4.207	0.00
Roman Catholic	-0.097	0.200	-0.484	0.63
Other religion	-0.710	0.193	-3.685	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	-0.258	0.157	-1.641	0.10
Single	-0.038	0.176	-0.218	0.83
Separated & Widowed	-0.800	0.205	-3.895	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.492	0.117	4.189	0.00
Fair	0.134	0.182	0.736	0.46
Bad_ Very bad	0.100	0.350	0.285	0.78
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.367	0.155	-2.365	0.02
A little	-0.812	0.166	-4.884	0.00
A lot	-1.096	0.248	-4.417	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.096	0.154	-0.620	0.54
Higher grade	-0.463	0.163	-2.844	0.00
Standard School Grade	-0.436	0.165	-2.647	0.01
Not qualifications	-1.093	0.252	-4.331	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.207	0.188	1.103	0.27
IIIN Skilled non-manual	-0.409	0.248	-1.645	0.10
IIIM Skilled manual	0.010	0.226	0.042	0.97
IV Semi-skilled manual	-0.431	0.255	-1.688	0.09
V Unskilled manual & other	-0.038	0.303	-0.124	0.90
Employment Status (neconacb)				
In employment	-	-	-	-
ILO unemployed & Inactive	-0.254	0.154	-1.645	0.10
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.548	0.114	4.827	0.00
Light smokers	0.580	0.273	2.121	0.03
Moderate smokers	1.180	0.223	5.287	0.00
Heavy smokers	1.330	0.311	4.271	0.00

TABLE I.41 COEFFICIENTS FROM CROSS-VALIDATION IN THE SMOTE TRAINING RURAL DATASET k = 10 / or = 1 (smk10r100_dwr)

Variable-category	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.375	0.301	-1.245	0.21
Survey year (syear)				
2017	-	-	-	-
2018	0.304	0.155	1.967	0.05
2019	0.578	0.147	3.933	0.00
2021	0.640	0.149	4.282	0.00
Sex (sex)				
Male	-	-	-	-
Female	-0.717	0.098	-7.315	0.00
Age (ag16g10)				
25-34	-	-	-	-
35-44	0.767	0.190	4.046	0.00
45-54	0.927	0.185	5.021	0.00
55-64	1.048	0.192	5.459	0.00
Birthplace (birthpla3)				
Scotland	-	-	-	-
Rest UK	-0.970	0.239	-4.050	0.00
Elsewhere	-2.441	0.487	-5.014	0.00
Ethnic (ethnic05)				
White: Scotland	-	-	-	-
White: Rest UK	0.677	0.244	2.774	0.01
Other	-0.222	0.440	-0.505	0.61
Religion (religi04)				
None	-	-	-	-
Church of Scotland	-0.549	0.130	-4.216	0.00
Roman Catholic	-0.065	0.195	-0.333	0.74
Other religion	-0.768	0.182	-4.216	0.00
Marital status (maritalg)				
Married / civil partnership	-	-	-	-
Living as married	-0.273	0.148	-1.845	0.07
Single	-0.161	0.169	-0.955	0.34
Separated & Widowed	-0.916	0.193	-4.753	0.00
General health (genhelf2)				
Very good	-	-	-	-
Good	0.345	0.111	3.110	0.00
Fair	0.093	0.168	0.555	0.58
Bad_ Very bad	-0.153	0.345	-0.444	0.66
LTC (limitac_h)				
Not limitation	-	-	-	-
Not at all	-0.647	0.153	-4.222	0.00
A little	-0.593	0.151	-3.924	0.00
A lot	-1.181	0.225	-5.243	0.00
Education (hedqul08)				
Degree	-	-	-	-
HNC/D	-0.232	0.147	-1.577	0.11
Higher grade	-0.538	0.153	-3.515	0.00
Standard School Grade	-0.458	0.156	-2.935	0.00
Not qualifications	-1.104	0.244	-4.527	0.00
Social class (schrpg7)				
I Professional	-	-	-	-
II Managerial technical	0.233	0.178	1.314	0.19
IIIN Skilled non-manual	-0.350	0.232	-1.508	0.13
IIIM Skilled manual	-0.177	0.218	-0.812	0.42
IV Semi-skilled manual	-0.200	0.237	-0.844	0.40
V Unskilled manual & other	-0.054	0.294	-0.182	0.86
Employment Status (neconacb)				
In employment	-	-	-	-
ILO unemployed & Inactive	-0.239	0.143	-1.667	0.10
SIMD quintile (SIMD20_RPa)				
Least deprived				
4th	0.094	0.175	0.538	0.59
3rd	0.050	0.180	0.280	0.78
2nd	-0.165	0.232	-0.711	0.48
Most deprived	-1.049	0.483	-2.175	0.03
Smoking (cig)				
Never smoked	-	-	-	-
Ex-smoker	0.522	0.107	4.897	0.00
Light smokers	0.794	0.258	3.076	0.00
Moderate smokers	0.821	0.225	3.650	0.00
Heavy smokers	1.055	0.312	3.380	0.00

TABLE I.42 ESTIMATES AND OR WITH 95% CI OF THE FINAL MODEL FOR THE URBAN & RURAL DATASET

Variable	Estimate	Std..Error	Z.value	P.value	OR	CI.Lower	CI.Upper
(Intercept)	0.153	0.109	1.408	0.16	1.165	0.942	1.441
Survey year (syear)							
2017	-	-	-	-	-	-	-
2018	0.272	0.066	4.105	0.00	1.313	1.153	1.495
2019	0.485	0.064	7.539	0.00	1.624	1.432	1.843
2021	0.762	0.065	11.745	0.00	2.142	1.887	2.433
Age (ag16g10)							
25-34							
35-44	0.079	0.070	1.123	0.26	1.082	0.943	1.241
45-54	0.567	0.068	8.278	0.00	1.763	1.541	2.016
55-64	0.800	0.072	11.142	0.00	2.225	1.933	2.561
Sex (sex)							
Male	-	-	-	-	-	-	-
Female	-0.784	0.042	-18.473	0.00	0.456	0.420	0.496
Ethnicity (ethnic05)							
White: Scotland	-	-	-	-	-	-	-
White: Rest UK	-0.328	0.064	-5.147	0.00	0.720	0.636	0.816
Other	-0.978	0.090	-10.816	0.00	0.376	0.315	0.449
Religion (religi04)							
None	-	-	-	-	-	-	-
Church of Scotland	-0.097	0.055	-1.777	0.08	0.907	0.815	1.010
Roman Catholic	-0.255	0.068	-3.749	0.00	0.775	0.678	0.885
Other religion	-0.920	0.088	-10.422	0.00	0.399	0.335	0.474
Marital status (maritalg)							
Married	-	-	-	-	-	-	-
As Married	-0.072	0.063	-1.131	0.26	0.931	0.822	1.054
Single	-0.178	0.064	-2.765	0.01	0.837	0.738	0.950
Separated & Widowed	-0.323	0.073	-4.403	0.00	0.724	0.627	0.836
General health (genhelf2)							
Very good	-	-	-	-	-	-	-
Good	0.111	0.049	2.259	0.02	1.117	1.015	1.230
Fair	-0.131	0.077	-1.695	0.09	0.878	0.755	1.021
Bad_ Very bad	0.275	0.128	2.154	0.03	1.316	1.025	1.690
LTC (limitac_h)							
Not limitation	-	-	-	-	-	-	-
Not at all	-0.226	0.064	-3.549	0.00	0.798	0.705	0.904
A little	-0.458	0.067	-6.877	0.00	0.632	0.555	0.721
A lot	-0.617	0.095	-6.468	0.00	0.540	0.448	0.651
Activity level (adt10gptw)							
Meet recommendations	-	-	-	-	-	-	-
Low activity	-0.485	0.068	-7.181	0.00	0.616	0.539	0.703
Very low activity	-0.542	0.075	-7.271	0.00	0.582	0.503	0.673
Smoking (cig)							
Never smoked	-	-	-	-	-	-	-
Ex-smoker	0.683	0.047	14.437	0.00	1.979	1.804	2.171
Light smokers	0.718	0.099	7.224	0.00	2.051	1.688	2.492
Moderate smokers	0.694	0.090	7.748	0.00	2.002	1.680	2.387
Heavy smokers	0.939	0.117	8.047	0.00	2.558	2.035	3.215
Educational level (hedqul08)							
Degree	-	-	-	-	-	-	-
HNC/D	-0.196	0.067	-2.940	0.00	0.822	0.722	0.937
Higher grade	-0.333	0.068	-4.872	0.00	0.717	0.627	0.820
Standard School Grade	-0.319	0.067	-4.738	0.00	0.727	0.637	0.829
Not qualifications	-0.321	0.095	-3.374	0.00	0.726	0.602	0.874
Social class (schrpg7)							
I Professional	-	-	-	-	-	-	-
II Managerial technical	0.391	0.074	5.297	0.00	1.479	1.279	1.709
IIIN Skilled non-manual	-0.140	0.093	-1.509	0.13	0.869	0.724	1.043
IIIM Skilled manual	0.004	0.091	0.047	0.96	1.004	0.841	1.200
IV Semi-skilled manual	-0.074	0.100	-0.742	0.46	0.928	0.763	1.130
V Unskilled manual & other	-0.015	0.127	-0.122	0.90	0.985	0.768	1.263
Economy activity (neconacb)							
In employment	-	-	-	-	-	-	-
ILO unemployed & Inactive	-0.285	0.062	-4.606	0.00	0.752	0.666	0.849

SIMD quintile (SIMD20_RPa)							
Least deprived	-	-	-	-	-	-	-
4th	-0.256	0.063	-4.079	0.00	0.774	0.685	0.876
3rd	-0.379	0.067	-5.653	0.00	0.685	0.600	0.781
2nd	-0.527	0.068	-7.704	0.00	0.590	0.516	0.675
Most deprived	-0.448	0.077	-5.832	0.00	0.639	0.550	0.743
Urban-rural class (urbrur_all)							
Urban	-	-	-	-	-	-	-
Rural	-0.312	0.058	-5.388	0.00	0.732	0.654	0.820

TABLE I.43 ESTIMATES AND OR WITH 95% CI OF THE FINAL MODEL FOR THE URBAN DATASET

Variable	Estimate	Std..Error	Z.value	P.value	OR	CI.Lower	CI.Upper
(Intercept)	0.199	0.123	1.620	0.11	1.220	0.959	1.553
Survey year (syear)							
2017	-	-	-	-	-	-	-
2018	0.052	0.073	0.712	0.48	1.053	0.913	1.214
2019	0.410	0.070	5.813	0.00	1.506	1.312	1.729
2021	0.628	0.071	8.820	0.00	1.875	1.630	2.156
Age (ag16g10)							
25-34	-	-	-	-	-	-	-
35-44	0.269	0.078	3.442	0.00	1.309	1.123	1.526
45-54	0.710	0.077	9.236	0.00	2.034	1.750	2.365
55-64	0.912	0.081	11.305	0.00	2.490	2.125	2.916
Sex (sex)							
Male	-	-	-	-	-	-	-
Female	-0.858	0.047	-18.094	0.00	0.424	0.386	0.465
Birthplace (birthpla3)							
Scotland	-	-	-	-	-	-	-
Rest UK	0.149	0.124	1.199	0.23	1.160	0.910	1.479
Elsewhere	-0.198	0.160	-1.240	0.22	0.820	0.600	1.122
Ethnicity (ethnic05)							
White: Scotland	-	-	-	-	-	-	-
White: Rest UK	-0.574	0.122	-4.713	0.00	0.563	0.444	0.715
Other	-1.008	0.167	-6.051	0.00	0.365	0.263	0.506
Religion (religi04)							
None	-	-	-	-	-	-	-
Church of Scotland	-0.118	0.062	-1.914	0.06	0.888	0.787	1.003
Roman Catholic	-0.269	0.074	-3.635	0.00	0.764	0.661	0.884
Other religion	-0.741	0.098	-7.530	0.00	0.477	0.393	0.578
Marital status (maritalg)							
Married	-	-	-	-	-	-	-
As Married	0.078	0.072	1.085	0.28	1.081	0.939	1.246
Single	-0.106	0.070	-1.521	0.13	0.900	0.785	1.031
Separated & Widowed	-0.173	0.080	-2.158	0.03	0.841	0.719	0.984
General health (genhelf2)							
Very good	-	-	-	-	-	-	-
Good	-0.095	0.056	-1.706	0.09	0.909	0.815	1.014
Fair	-0.308	0.084	-3.665	0.00	0.735	0.623	0.866
Bad_ Very bad	0.085	0.133	0.644	0.52	1.089	0.840	1.413
LTC (limitac_h)							
Not limitation	-	-	-	-	-	-	-
Not at all	-0.269	0.074	-3.642	0.00	0.764	0.661	0.883
A little	-0.417	0.075	-5.548	0.00	0.659	0.569	0.764
A lot	-0.529	0.101	-5.226	0.00	0.589	0.483	0.719
Activity level (adt10gptw)							
Meet recommendations	-	-	-	-	-	-	-
Low activity	-0.449	0.074	-6.052	0.00	0.638	0.552	0.738
Very low activity	-0.552	0.082	-6.725		0.576	0.490	0.676
				0.00			
Smoking (cig)							
Never smoked	-	-	-	-	-	-	-
Ex-smoker	0.723	0.054	13.500	0.00	2.060	1.855	2.288
Light smokers	0.847	0.104	8.123	0.00	2.333	1.902	2.862
Moderate smokers	0.655	0.097	6.719	0.00	1.925	1.590	2.330
Heavy smokers	1.011	0.125	8.072	0.00	2.748	2.150	3.512
Educational level (hedqul08)							
Degree	-	-	-	-	-	-	-
HNC/D	-0.263	0.078	-3.366	0.00	0.769	0.659	0.896
Higher grade	-0.230	0.075	-3.086	0.00	0.794	0.686	0.919
Standard School Grade	-0.362	0.105	-3.453	0.00	0.696	0.567	0.855
Not qualifications	-0.263	0.078	-3.366	0.00	0.769	0.659	0.896
Social class (schrpg7)							
I Professional	-	-	-	-	-	-	-
II Managerial technical	0.257	0.084	3.041	0.00	1.293	1.096	1.525
IIIN Skilled non-manual	-0.149	0.104	-1.430	0.15	0.861	0.702	1.057
IIIM Skilled manual	0.064	0.103	0.626	0.53	1.066	0.872	1.304
IV Semi-skilled manual	0.036	0.112	0.319	0.75	1.036	0.833	1.290
V Unskilled manual & other	-0.032	0.141	-0.230	0.82	0.968	0.734	1.276

SIMD quintile (SIMD20_RPa)							
Least deprived	-	-	-	-	-	-	-
4th	-0.292	0.071	-4.125	0.00	0.747	0.650	0.858
3rd	-0.425	0.076	-5.605	0.00	0.653	0.563	0.758
2nd	-0.533	0.072	-7.352	0.00	0.587	0.509	0.677
Most deprived	-0.536	0.082	-6.561	0.00	0.585	0.499	0.687

TABLE I.44 ESTIMATES AND OR WITH 95% CI OF THE FINAL MODEL FOR THE RURAL DATASET

Variable	Estimate	Std..Error	Z.value	P.value	OR	CI.Lower	CI.Upper
(Intercept)	-0.371	0.265	-1.403	0.16	0.690	0.411	1.159
Survey year (syear)							
2017	-	-	-	-	-	-	-
2018	0.301	0.150	2.005	0.05	1.352	1.007	1.814
2019	0.392	0.147	2.665	0.01	1.480	1.109	1.974
2021	0.720	0.146	4.916	0.00	2.054	1.542	2.737
Sex (sex)							
Male							
Female	-0.759	0.097	-7.803	0.00	0.468	0.387	0.566
Age (ag16g10)							
25-34							
35-44	0.614	0.185	3.316	0.00	1.848	1.285	2.656
45-54	0.918	0.177	5.170	0.00	2.503	1.768	3.544
55-64	0.892	0.184	4.847	0.00	2.439	1.701	3.498
Birthplace (birthpla3)							
Scotland							
Rest UK	-0.857	0.227	-3.772	0.00	0.424	0.272	0.662
Elsewhere	-1.966	0.422	-4.656	0.00	0.140	0.061	0.320
Ethnicity (ethnic05)							
White: Scotland							
White: Rest UK	0.628	0.232	2.712	0.01	1.874	1.190	2.951
Other	-0.137	0.401	-0.342	0.73	0.872	0.397	1.915
Religion (religi04)							
None							
Church of Scotland	-0.553	0.128	-4.316	0.00	0.575	0.448	0.740
Roman Catholic	-0.486	0.211	-2.308	0.02	0.615	0.407	0.929
Other religion	-0.770	0.176	-4.376	0.00	0.463	0.328	0.654
Marital status (maritalg)							
Married & As Married							
Single	-0.141	0.164	-0.863	0.39	0.868	0.630	1.197
Separated & Widowed	-0.912	0.192	-4.739	0.00	0.402	0.275	0.586
General health (genhelf2)							
Very good							
Good	0.489	0.110	4.454	0.00	1.630	1.315	2.022
Fair	0.109	0.174	0.626	0.53	1.115	0.793	1.567
Bad_ Very bad	0.684	0.347	1.972	0.05	1.983	1.004	3.915
LTC (limitac_h)							
Not limitation							
Not at all	-0.399	0.147	-2.709	0.01	0.671	0.503	0.896
A little	-0.572	0.147	-3.882	0.00	0.564	0.423	0.753
A lot	-1.447	0.256	-5.643	0.00	0.235	0.142	0.389
Educational levels (hedqul08)							
Degree & HNC/D							
Higher grade	-0.423	0.115	-3.665	0.00	0.655	0.522	0.821
Standard School Grade	-0.674	0.151	-4.470	0.00	0.510	0.379	0.685
Not qualifications	-1.006	0.227	-4.434	0.00	0.366	0.234	0.570
SIMD quintile (SIMD20_RPa)							
Least deprived							
4th	0.142	0.172	0.825	0.41	1.152	0.823	1.614
3rd	0.187	0.174	1.072	0.28	1.205	0.857	1.695
2nd	-0.259	0.230	-1.126	0.26	0.772	0.491	1.212
Most deprived	-1.249	0.535	-2.335	0.02	0.287	0.100	0.818
Physical activity level (adt10gptw)							
Meet recommendations							
Low activity	-0.064	0.151	-0.424	0.67	0.938	0.697	1.261
Very low activity	-0.466	0.190	-2.456	0.01	0.628	0.433	0.910
Smoking (cig)							
Never smoked							
Ex-smoker	0.466	0.107	4.355	0.00	1.593	1.292	1.965
Light smokers	1.011	0.247	4.092	0.00	2.747	1.693	4.458
Moderate smokers	1.099	0.214	5.142	0.00	3.000	1.974	4.561
Heavy smokers	1.242	0.301	4.128	0.00	3.461	1.920	6.242

ANNEXES

ANNEX A. COPY OF EMAIL RECEIVED FROM UK DATA SERVICE

Gmail Loliett Valdés <valdes.loliett@gmail.com>

[JIRA] [QTHELP-62958] Scottish Health Survey

Sadiq Rahman <replyto@ukdataservice.ac.uk> 30 May 2024 at 11:34
Reply-To: replyto@ukdataservice.ac.uk
To: valdes.loliett@gmail.com

Dear Valdes,

Thank you for getting in touch and apologies for the delay in getting back to you.

We currently don't have the 2022 SHeS data, we are awaiting deposit of this data. We have asked the ScotCen who deposit this data with us to provide us a timeframe on when we can expect this, and we are waiting to hear back from them.

Best wishes,
Sadiq

Please ensure that the subject is left unaltered when replying to this message, for your response to be processed.

Sadiq Rahman

T +44(0) 1206 872143
E help@ukdataservice.ac.uk
W ukdataservice.ac.uk/help/get-in-touch

UK Data Service
UK Data Archive
University of Essex

Legal Disclaimer: Any views expressed by the sender of this message are not necessarily those of the UK Data Service or the UK Data Archive. This email and any files with it are confidential and intended solely for the use of the individual(s) or entity to whom they are addressed.

ANNEX B. CONVERSION USED FOR THE SHeS TO CONVERT VOLUMES OF ALCOHOL REPORTED IN THE SURVEY INTO UNITS

Alcohol unit conversion factors

Type of drink	Volume reported	Unit conversion factor
Normal strength beer, lager, stout, cider, shandy (less than 6% Alcohol By Volume (ABV)	Half pint	1.0
	Can or bottle	Amount in pints multiplied by 2.5
	Small can (size unknown)	1.5
	Large can / bottle (size unknown)	2.0
Strong beer, lager, stout, cider, shandy (6% ABV or more)	Half pint	2.0
	Can or bottle	Amount in pints multiplied by 4
	Small can (size unknown)	2.0
	Large can / bottle (size unknown)	3.0
Wine (including champagne and prosecco)	250ml glass	3.0
	175ml glass	2.0
	125ml glass	1.5
	750ml bottle	1.5 x 6
Sherry, vermouth and other fortified wines	Glass	1.0
Spirits	Glass (single measure)	1.0
Alcopops	Small can or bottle	1.5
	Large (700ml) bottle	3.5