A Study of Machine Learning Models in Predicting Young People and Adults Cigarette Smoking Status

CS910 Foundations of Data Analytics Project

2138473
University of Warwick, UK

ABSTRACT

Smoking has become one of the most detrimental causes in inferior health outcomes. This study seeks to explore imperative factors which might be predictive to smoking status among the population by analysing the large-scale health survey data. Using the 2019 Health Survey for England data, this study firstly identified trends and patterns regarding smoking status with socio-demographics, well-being, and substance-use/-exposure features. After that, using machine learning methods, 4 classification models has been built (Random Forest, Naive Bayes, Support Vector Machine and K-nearest Neighbors). Overall, 4 models achieve good accuracy (over 0.7) when using general representative features to predict sample smoking status. Finally, limitations and plausible applications of this study was concluded.

1 INTRODUCTION

Cigarette smoking is one of the primary causes in morbidity and mortality worldwide [1]. It has been estimated that, during 2017-2019, up to 191,900 deaths can be ascribed to smoking in England as reported by UK government [2]. Specifically regarding health issues, extensive empirical literature suggest that smoking frequently occurs in young and middle-aged population and is associate with increased probability in cardiovascular diseases at a later age, acute blood pressure elevation, and obesity. [3-5]. Aside from physical health risk factors, cigarette smoking was also suggested to link with socio-economic and mental health factors such as low financial status, low educational attainment, unemployment, childhood exposure to nicotine, parental and peer influence, depression and anxiety, female gender with eating disorders[6-14]. Also, these factors are thought to be enablers to aggravate smoking dependency, although some of the causal relationships remain unclear [15][16].

Research that investigate common factors associated with tobacco-use, in a sense, could provide the public health services with better understanding to predict and control for cigarette consumption. To date, a large body of study that investigates the nature of tobacco-use and its related causal factors adopted inferential and regression analysis, while this could be regarded as a methodological limitation for researchers. Therefore, to expand the understanding and application of the knowledge of complicated relationships between smoking and its possible causes, this exploratory study sought to employ predictive models established from machine learning (ML) algorithms.

CS910, PGT Project, 2021/22
2138473.

2 BACKGROUND

Since ML have been introduced in health informatics studies, there consequently emerges a growing body of empirical applications of ML algorithms on tobacco research[17][18]. According to a review from *Fu et al.*, Naive Bayes, K-nearest neighbors and Random Forest algorithms are frequently applied in ML studies using survey or clinical trial data [18].

More specifically, ML studies that achieve relative high performance (around 0.8 at accuracy or precision) vary in the selection of candidate features, but attributes such as socio-demographics, well-being, substance use are most commonly included [18-21]. Therefore, the current study adopted methods with regards to data processing and classifier implementing based on previous work and aims to attain high model performance.

3 KEY OBJECTIVES

Contribution Summary. The current study is a secondary analysis of 2019 Health Survey for England data, which was obtained from a large-scale, nationwide annual study that aims to evaluate and reflect the national health level. The key objective of the current enquiry are:

- To observe patterns of smoking prevalence among young to middle-aged people. Hence, it is then hypothesised that some noticeable patterns of tobacco usage could be identified in the population when considering attribute classes that were previous suggested to be prone to cigarette dependence.
- Implementing classification models to predict sample cigarette smoking status from the health data published by UK government.
- To evaluate the classification models and provide insights with tobacco-related health research.

4 METHOD

4.1 Software Tools

R and RStudio. RStudio is an intergrated development environment (IDE) for R, which is a powerful programming language for statistical data analysis and data visualisation. RStudio Version 1.4 including packages *"tidyverse", "ggplot2", "foreign", etc,.* were employed for data cleansing and graphing [22].

Weka. Waikato Environment for Knowledge Analysis (Weka) is an advanced, free-access software which incorporates of various predictive data modelling algorithms. Weka Version 3.8.5 was deployed to serve the purpose of building multi-class classification model in this study [23].

4.2 Data Preprocessing

THE DATA. The 2019 Health Survey for England (HSE) data were used for analysis. The datasets were accessed through UK data service using University of Warwick postgraduate student end user licence. In detail, HSE is a series of surveys specifically designated for the purpose of monitoring and identifying UK nationwide health trend. The large-scale annual survey was authorised by NHS Digital and proceeded by the Joint Health Surveys Unit of the National Centre for Social Research and the Department of Epidemiology and Public Health at University College London [24].

The initial dataset comprised of 10299 records, with 1841 encoded variables (file size 21Mb in total) retrieved from *(1) Computerised and paper-based self-administered questionnaires; (2) Clinical measurements; (3) Computer-assisted face-to-face interviews,* which were conducted from 01/01/2019 to 01/03/2020.

DATA CLEANING. Primarily, this study intentionally restricted the target sample to young people (Age: 16-19) and young to middle-aged adults (Age: 20-54). Thus using the *Variable Name = Age35g; label = Age, 3 year bands for 0-15, 5 year bands 16+*, this study excluded participants who were aged below 16 and above 55.

Furthermore, to simplify the dataset for further analysis, the unwarranted attributes in relation to cigarette smoking status, such as variables which were created for storing interview information (*e.g., Whether the participant mentioned smoking in the interview?*), other less-related demographics variables or with a considerable numbers of missing cases (*e.g., Household size, Diabetes history, Whether living with caregiver, etc,.*), and variables belong to the same characteristics but coded in different methods (*e.g., age in 5 year bands and age in 10 year bands*) were filtered and excluded from this study.

Consequently, a total of 16 attributes were retained from the initial dataset, which covered 5 core aspects including demographics characteristics (*Age, Gender, Ethnicity, Religion, Education Level, Occupation*), general well-being (*Self-assessed general health, Overall life satisfaction*), other health factors (*Body-Mass Index (BMI) group, Eating disorder, Blood pressure level*), alcohol consumption (*Risk group - alcohol units per week*) and smoking-related factors (*E-cigarette use, Childhood /Adulthood exposure to cigarette smokers, Current cigarette smoking status*).

Finally, since the original HSE dataset coded the smoking status attribute in to 4 classes: *(1) Current cigarette smoker; (2) Used to smoked cigarettes regularly; (3) Used to smoked cigarettes occasionally; (4) Never smoked cigarettes at all.* Therefore, to create a binary classes for the outcome feature in order to simplify the modelling process, the current study recoded the target feature into 2 levels: *(1) Current cigarette smoker or have smoking history; (2) Never smoked cigarettes at all.*

MISSING VALUES. In general, this study adopted 2 separate methods in handling missing values. For attributes whose missing records are less than 5% of total instances, the missing values were simply excluded for analysis. Nonetheless, there are certain imperative attributes which their total missing instances were weighing more than 5% and up to around 10% (*Childhood exposure to smokers, Eating disorder, Risk group - alcohol units per week*), the missing values were transformed to a new class named *"unknown"*

Finally, a total of 3670 instances with 16 attributes were retained from the initial dataset. The descriptive analysis about each data element will be displayed in the following sections.

4.3 Descriptive Analysis

To begin with observing and identifying patterns in the data, this study firstly looked at the descriptive statistics of each attribute. The visualisation and tidying process were proceeded in R and RStudio. In detail, the study calculated the number of total instances and its the proportion within the sample of each smoking status feature. Furthermore, the socio-demographics and general well-being attributes were plotted in an interactive way with smoking status, while the rest of the attributes were listed accordingly in a table in **Results** section.

4.4 Implementation of Classifier

The built-in Random Forest, Naive Bayes, Support Vector Machine (SVM) and K-nearest Neighbors (KNN) algorithms in Weka were employed for building classification models. The accuracy, area under receiver operating characteristic curve (AUC-ROC) and kappa statistics are extracted to evaluate the model performance.

Random Forest. Random Forest Classifier, in literal meaning, refers to forming a collection of decision trees (i.e., forest). The algorithm extracts random samples from training data and then develop individual decision tree models. Each tree in the random forest generates an independent class prediction. Then the class voted by the most learned trees will be returned as model's prediction.

Naive Bayes. Naive Bayes (NB) Classifier is an effective method for classification stems from the Bayes' theorem, which is also a widely-applied classification method when dealing with large dataset. Basically, Naive Bayes assume that for random value of a predictor X in the training set, its effect on a specific C with m classes $C_1, C_2, ..., C_m$ is independent. By evaluating the posterior probability $P(C_i|X)$, NB algorithm then get the class with maximum posterior probability to be the outcome of the prediction. The $P(C_i|X)$ is calculated as follows [25]:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)},$$

$$P(X|C_i)P(C_i) = \prod_{k=1}^{n} P(x_k|C_i),$$

where x_k is the value of X.

Support Vector Machine. SVM is a useful algorithm when dealing with classifying high dimensional data. In detail, it applies a non-linear mapping by reconstructing the training set into a new (higher) dimension, where it could find the linear best-fit hyperplane (i.e., decision boundary). Therefore, the hyperplane distinguishes the data from 2 classes with suitable non-linear mapping. A separating hyperplane is defined as:

$$\boldsymbol{W} \cdot \boldsymbol{X} + b = 0,$$

where $\boldsymbol{W} = (w_1, w_2, ..., w_n)$, is a weight vector, n stands for the number of features, the scalar b means bias. Although the training time of SVM, in relative terms, is slow comparing to other algorithms, yet SVM could promise high accuracy and less prone to overfitting problems [25].

K-nearest Neighbors. KNN is a non parametric classification method, where its input consists of the K closest training instances in a dataset, and the most representative class by these k nearest neighbours is considered the output class label. This study utilises Euclidean distances as index distances to find the nearest neighbors. The Euclidean distance is calculated as follows [25]:

$$d(i, j) = \sqrt{(i_1 - j_1)^2 + (i_2 - j_2)^2 + ... + (i_n - j_n)^2}$$

AUC-ROC. An ROC curve is a visualisation of classification model performance. In the curve, the True Positive Rate (TPR; also known as recall) is plotted against the False Positive Rate (FPR) at various classification thresholds. The TPR and FPR are defined as follows:

$$TPR = \frac{TP}{TP + FN}$$

,

$$FPR = \frac{FP}{FP + TN}$$

,

where TP,FP,TN and FN are numbers of true positives, false positives, true negatives and false negatives.

Hence, AUC-ROC denotes *"Area under the ROC Curve"*, which is a measure of the capability of a model to discriminate various classes and is often utilised to encapsulate the ROC curve. Normally, an AUC value equal to 0.5 indicates no discriminability, while AUC = 1 suggests the classifier exactly and marvellously distinguish distinct classes. Thus, higher AUC values demonstrating better performance in discerning classes [26].

Cohen's Kappa. Cohen's Kappa statistic is an effective metric in evaluating model performance when facing imbalanced classes issues. Cohen's Kappa is defined as:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

,

where $Pr(a)$ indicates the observed agreement and $Pr(e)$ illustrates the chance agreement. In simple terms, Cohen's Kappa suggest how well the extent to which a model performs by random estimation based on the class frequency. The κ value is always less than or equal to 1. Previous literature suggest than a κ less than 0 demonstrates no agreement, 0-0.20 is regarded as slight agreement, 0.21-0.40 as acceptable, 0.41-0.60 as average, and 0.61-1 indicating good to excellent [27].

The Naive Bayes, Random Forest and SVM algorithms were performed in Weka using default parameters. Additionally, KNN (IBk) was performed with setting K to 7. Specifically, this study randomly selected 66% the total instances as training set, while the remaining 34% were used for testing the model. The model performance metrics are listed in tables in the next section.

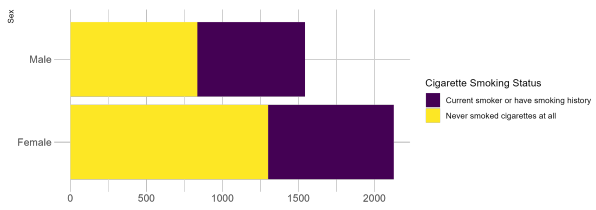

Figure 1: Cigarette smoking status distribution with Sex

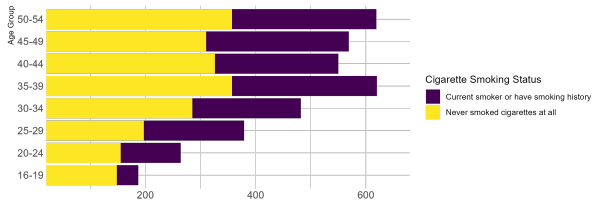

Figure 2: Cigarette smoking status distribution with Age

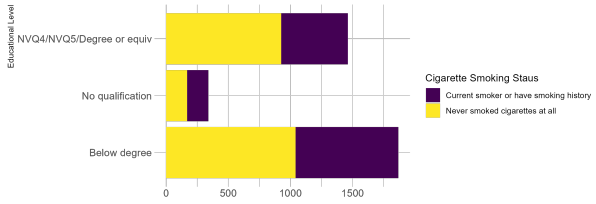

Figure 3: Cigarette smoking status distribution with Educational level

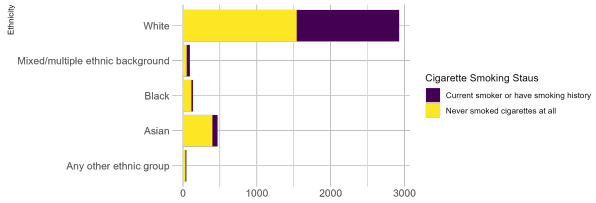

Figure 4: Cigarette smoking status distribution with Ethnicity

Variable	Current Cigarette Smoker or have smoking history N = 1535 (42%)	Non-smoker N = 2135 (58%)
Doctor diagnosed high blood pressure (excluding pregnant)	Yes: N = 188 (5.12%); No: N = 1347 (36.70%)	Yes: N = 180 (4.9%); No: N = 1955 (53.3%)
Parents ever smoked regularly when a child	Yes, both: N = 476 (12.97%); Yes, one: N = 545 (14.85%); No, Neither: N = 489 (13.32%); Unknown: N = 25 (0.69%)	Yes, both: N = 349 (9.51%); Yes, one: N = 635 (17.30%); No, Neither: N = 1040 (2.32%); Unknown: N = 111 (28.34%)
Any adult self-reported exposure to other people's smoke, 16+	Yes: N = 716 (19.5%); No: N = 819 (22.3%)	Yes: N = 548 (14.9%); No: N = 1587 (43.2%)
Ever used e-cigarette or vaping device	Yes: N = 645 (17.6%); No: N = 890 (24.3%)	Yes: N = 27 (0.7%); No: N = 2108 (57.4%)
Alcohol units per week - risk groups (2016 guidelines)	Non drinker/ not in last 12 months: N = 17 (0.46%); Lower risk (up to 14 units): N = 892 (23.76%); Increased risk (14-50/14-35): N = 336 (9.16%); Higher risk (more than 50/35): N = 119 (3.24%); Unknown: N = 17 (0.46%)	Non drinker/ not in last 12 months: N = 523 (14.251%); Lower risk (up to 14 units): N = 1272 (34.659%); Increased risk (14-50/14-35): N = 273 N = 523 (7.439%); Higher risk (more than 50/35): N = 30 (0.871%); Unknown: N = 37 (1.008%)
BMI grouped combining obese and morbidly obese	Underweight: N = 232 (6.32%); Normal: N = 451 (12.27%); Overweight: N = 476 (12.97%); Obese: N = 376 (10.25%)	Underweight: N = 338 (9.21%); Normal: N = 693 (18.88%); Overweight: N = 594 (16.19%); Obese: N = 510 (13.90%)
SCOFF Score grouped 2+ with significant impact	Lower score and/or no impact: N = 1435 (39.101%); Score of 2+ and significant impact: N = 83 (2.262%); Unknown: N = 17 (0.463%)	Lower score and/or no impact: N = 2002 (54.55%); Score of 2+ and significant impact: N = 115 (3.13%); Unknown: N = 18 (0.14%)

Table 1: A holistic display of variables 9-16 regarding specific information and categorical value levels in the finalised dataset for analysis. *Note: The SCOFFIMP scores were deduced from the SCOFF Questionnaire. This measurement was originated by researchers at St George's Hospital Medical School, which is a widely-used tool with high validity and reliability in identifying eating disorders [28].

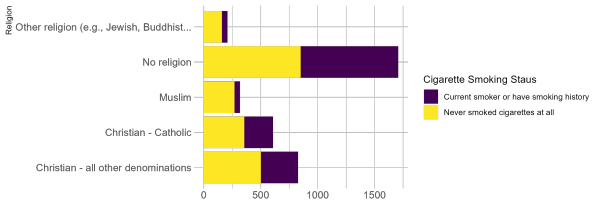

Figure 5: Cigarette smoking status distribution with Religion

Algorithm	Accuracy	Cohen's Kappa
Random Forest	75.4%	0.48
Naive Bayes	74.9%	0.47
SVM	77.1%	0.50
KNN	74.3%	0.44

Table 2: Accuracy and Kappa score for each classification model

5 RESULTS AND DISCUSSION

Descriptive Summary. Overall, 42% of the total sample (N = 1535; mean age = 38.59) are current or past cigarette smokers, while 58% (N = 2135, mean age = 37.69) reported never smoked cigarettes

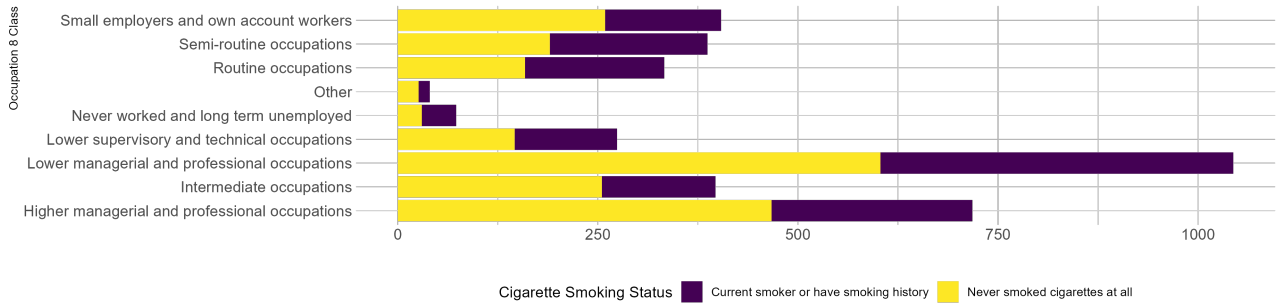

Figure 6: Cigarette smoking status distribution with Occupation class

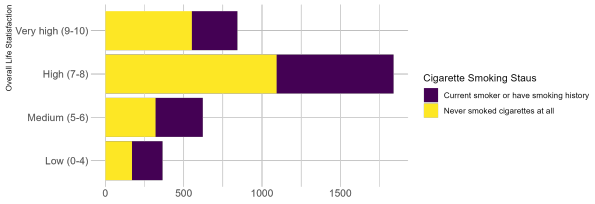

Figure 7: Cigarette smoking status distribution with Life satisfaction

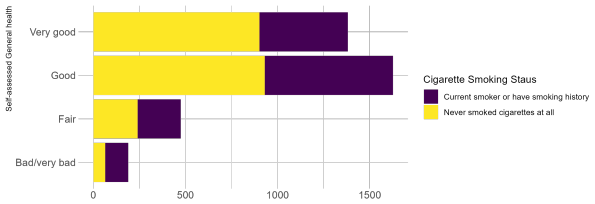

Figure 8: Cigarette smoking status distribution with Self-assessed well-being

Algorithm	AUC-ROC
Random Forest	Weighted Avg.= 0.813
Naive Bayes	Weighted Avg.= 0.823
SVM	Weighted Avg.= 0.734
KNN	Weighted Avg.= 0.804

Table 3: AUC of each classification model.

at all. Figure 1-8 demonstrate the distributions of demographic variables and general well-being factors. Speaking of gender, 58% of the total instances are female, while 42% are male, and there appears to be more female non-smokers (35.42%) than male (22.75%) in proportion. Most of the current or past smokers were White (N = 518, 14.114%), and Asian are relatively most likely to be non-smokers (N = 398, 10.845%) in comparison to other ethnicity. Nearly half of the sample (46.5%) reported as no religious belief, and this group has the largest portion of smokers (N = 857, 23.35%) comparing to other religious groups.

To go further, it has been shown that among all attributes, non-smoking status is relatively common in sample who were aged between 16 and 19, gender as female, in Muslim and Other religious group and in Asian ethnicity groups, had higher education attainment (NVQ4/NVQ5/Degree or equivalent) and reported very high overall life satisfaction. Additionally, Table 1 illustrates the descriptive representations of other attributes. Notably, we can see a relatively strong prevalence for non-smokers in relation to normal blood pressure level, no childhood/adulthood exposure to smokers, no e-cigarette use history and lower-risk- to non-drinkers. Nonetheless, as opposed to the initial hypothesis, for both current smokers and past smokers, there seems to be few patent patterns to be recognised, except for the e-cigarette use factor, it might be due to the imbalanced class where there are fewer current pr past cigarette use records obtained in the data.

Classification Model Summary. Table 2 and 3 are summaries of classification accuracy and other performance metrics. In general, models built by Random Forest (RF), Naive Bayes (NB), SVM and KNN classifier algorithms achieve good accuracy, 75.4%, 74.9%, 77.1% and 74.3% respectively, in predicting cigarette smoking status. As we can see, SVM outperforms other 3 classifiers when concerning the prediction accuracy. Furthermore, when considering the AUC-ROC values, all three model did good job in showing discriminability (RF: 0.813 vs NB: 0.823 vs SVM: 0.734 vs KNN: 0.804). Whereas only SVM gives the AUC value smaller than 0.8, indicating a less good discriminability between classes. Furthermore, for the Cohen's Kappa, the Kappa scores for all 4 classifiers fall in the mid-range (RF: 0.48 vs NB: 0.47 vs SVM: 0.5 vs KNN: 0.44) indicating a reasonable agreement

In general, classification models show reasonable accuracy and AUC-ROC in classifying smoking status. Among 4 models, SVM slightly outperforms RF, NB and KNN regarding accuracy and Cohen's Kappa but with an average AUC-ROC value. Indeed, the accuracy of at least 74.3% indicates there exists some substantial

internal relationships between the predictor variables and smoking status.

Merits and Limitations. The current inquiry introduced ML methods and using general features in predicting participants' current smoking status. The RF, NB, SVM and KNN models, altogether, attained comparable good accuracy and other performance metrics when predicting the tobacco usage class within the data. Consequently, it implies that these factors are, to some extent, obtaining substantial relations with smoking intentions, and this should be aware by the public sections.

However, considering the feature selection, the current study did not include ML methods to adjust the feature selection process. Instead, the selection was based on previous literature, which could be quite intuitive and biased. Therefore, future work should seek to systematically improve this procedure. Moreover, it can be seen that the current inquiry includes fewer attributes that have possible substantial relationship with smoking status, future study might sought to improve in these aspects. Previous research often includes more features which illustrates mood state, caffeine consumption, etc, [18].

Furthermore, this study only deploy the data from 2019 HSE survey data, which contains limited records (N =3670) after data cleaning. Thus, the finalised dataset might not be representative enough in training a model which could enough fit the real world setting. In addition to using survey data only, there exists some ML studies using external data such as tobacco related post on Twitter, Facebook, etc, [18]. Future studies might want to explore in this aspect as well.

6 CONCLUSION

The current study offers an intuitive comprehension in predictive health information analysis among sample with distinct smoking status. Specifically, this paper presents the comparison of results that have been achieved by applying different machine learning techniques for the prediction of smoking status. In particular, this study analysed the 2019 HSE dataset, which contains nationwide health records.

From the original dataset, the study select sample who obtained records that demonstrating their smoking status in 2 classes: Current or past smoker (quit smoking) and Non-smoker. In the data cleaning and feature selection process, this study choose the socio-demographics, general and specific well-being, substance-use/ -exposure related factors as features according to previous literature and firstly identified the patterns using descriptive statistical results. After that, current study build Random Forest, Naive Bayes, SVM and KNN classifiers to fit classification models to predict smoking status, and then evaluate the models accordingly.

Taken to all, the 4 models achieve good accuracy on average when using general factors as predictors. Nonetheless, as the current study did not use ML techniques to control for feature selection process, this might in turn, results in a loss in model performance. Therefore, future works could expand the current framework by perfecting the feature selection process with various ML techniques in order to achieve greater accuracy.

Applying ML methods to health informatics research is indeed, offers the public health section a high-efficient and low-cost way in dealing with complex health problems. Recognising and understanding these smoking-predictive factors and their underlying relationships is vital to decision and policy makers, such that they could develop more suitable strategies to the lay public and business co-operations in lowering the health burden in the society.

REFERENCES

1. World Health Organization. WHO global report on trends in prevalence of tobacco use 2000–2025. 2nd ed. World Health Organization (Geneva); 2019,

2. Public Health England. Local tobacco control profiles for England: short statistical commentary, July 2021. Public Health England (UK); 2021,

3. Jeffrey B. Lakier, Smoking and cardiovascular disease, The American Journal of Medicine, Volume 93, Issue 1, Supplement 1, 1992, Pages S8-S12, ISSN 0002-9343, https://doi.org/10.1016/0002-9343(92)90620-Q.

4. Courtemanche, C., Tchernis, R., Ukert, B.. (2018). The effect of smoking on obesity: Evidence from a randomized trial. Journal of Health Economics, 57, 31–44. https://doi.org/10.1016/j.jhealeco.2017.10.006,

5. Primatesta, P., Falaschetti, E., Gupta, S., Marmot, M. G., Poulter, N. R.. (2001). Association Between Smoking and Blood Pressure. Hypertension, 37(2), 187–193. https://doi.org/10.1161/01.hyp.37.2.187,

6. Cancer Council New South Wales, Addressing smoking in community service organisations: a policy toolkit, 2008, Cancer Council NSW: Sydney.

7. Gage, S. H., Sallis, H. M., Lassi, G., Wootton, R. E., Mokrysz, C., Davey Smith, G., Munafò, M. R.. (2020). Does smoking cause lower educational attainment and general cognitive ability? Triangulation of causal evidence using multiple study designs. Psychological Medicine, 1–9. https://doi.org/10.1017/s0033291720003402

8. Macleod, J., Hickman, M, Bowen, E, Alati, R, Tilling, K, Smith, G Parental drug use, early adversities, later childhood problems and children's use of tobacco and alcohol at age 10: birth cohort study. Addiction., 2008. 103(10): p. 1731-43.

9. Siahpush, M., Borland, R., Scollo, M., Smoking and financial stress. Tobacco Control, 2003. 12: p. 60-66.

10. Cougle, J., Zvolensky, M, Fitch, K, Sachs-Ericsson, N, The role of comorbidity in explaining the associations between anxiety disorders and smoking. Nicotine Tobacco Research, 2010. 12(4): p. 355-64.

11. Richter, M., Vereecken, C, Boyce, W, Maes, L, Gabhainn, S, Currie, C, Parental occupation, family affluence and adolescent health behaviour in 28 countries. International Journal of Public Health, 2009. 54(4): p. 203-12.

12. John, R., Cheney, MK, Azad, MR, , Point-of-sale marketing of tobacco products: taking advantage of the socially disadvantaged? J Health Care Poor Underserved, 2009. 20: p. 489-506.

13. Apollonio, D., Malone, RE Marketing to the marginalised: tobacco industry targeting of the homeless and mentally ill. Tob Control 2005. 14: p. 409-15.

14. Smith, D., Leggat, P, Tobacco smoking by occupation in Australia: results from the 2004 to 2005 National Health Survey. Journal of Occupational and Environmental Medicine, 2007. 49(4): p. 437-45.

15. Anzengruber D, Klump KL, Thornton L, Brandt H, Crawford S, Fichter MM, Halmi KA, Johnson C, Kaplan AS, LaVia M, Mitchell J, Strober M, Woodside DB, Rotondo A, Berrettini WH, Kaye WH, Bulik CM. Smoking in eating disorders. Eat Behav. 2006 Nov;7(4):291-9. doi: 10.1016/j.eatbeh.2006.06.005. Epub 2006 Jun 27. PMID: 17056404.

16. The Department of Health of Australian Government, SMOKING DISADVANTAGE EVIDENCE BRIEF, The Department of Health of Australian Government (Australia), 2013,

17. Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., Yang, G. Z. (2016). Deep learning for health informatics. IEEE journal of biomedical and health informatics, 21(1), 4-21.

18. Fu R, Kundu A, Mitsakakis N, et alMachine learning applications in tobacco research: a scoping reviewTobacco Control Published Online First: 27 August 2021. doi: 10.1136/tobaccocontrol-2020-056438

19. Davagdorj K, Lee JS, Park KH. A machine-learning approach for predicting success in smoking cessation intervention. 2019 IEEE 10th International Conference on Awareness Science and Technology (ICAST), 2019:1–6.

20. Dumortier A, Beckjord E, Shiffman S, et al. Classifying smoking urges via machine learning. Comput Methods Programs Biomed 2016;137:203–13.

21. Kim N, McCarthy DE, Loh W-Y, et al. Predictors of adherence to nicotine replacement therapy: machine learning evidence that perceived need predicts medication use. Drug Alcohol Depend 2019;205:107668.

22. RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

23. Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

24. NatCen Social Research, University College London, Department of Epidemiology and Public Health. (2021). Health Survey for England, 2019. [data collection]. UK Data Service. SN: 8860, DOI: 10.5255/UKDA-SN-8860-1

25. Han, J., Kamber, M. (2001). Data mining: Concepts and techniques. San Francisco: Morgan Kaufmann Publishers.

26. J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 1982.

27. McHugh M. L. (2012). Interrater reliability: the kappa statistic. Biochemia medica, 22(3), 276–282.

28. Hill, L. S., Reid, F., Morgan, J. F., Lacey, J. H. (2010). SCOFF, the development of an eating disorder screening questionnaire. International journal of eating disorders, 43(4), 344-351.