e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

EXPLORATORY DATA MINING ON LOCAL AUTHORITIES' HEALTH PROFILES IN ENGLAND

Gilbert Umemba*1, Dr. Charith Silva*2, Prof. Mo Saraee*3

*1,2,3School Of Computing, Science, And Engineering, University Of Salford, Greater Manchester, United Kingdom.

ABSTRACT

This project seeks to apply exploratory data mining techniques to health profile data with the aim of extracting advanced analytic insights that would aid the understanding of public health issues and provide evidence-based literacy that will be utilized by local authorities in health policymaking. Clustering techniques were applied to the dataset to explore the possibility of grouping the local authorities into clusters based on their similarities with respect to the data attributes. Association rule mining was also performed to extract strong association rules showing dependency and correlation among data attributes. Trend analysis to investigate the trend in smoking profiles was implemented on the tobacco profile attributes. Our exploratory mining showed the existence of meaningful clusters in the dataset. The k-means method performed better than the hierarchical agglomerative technique with the formation of 3 well-separated clusters with good **si** width values from the silhouette plot. Strong association rules were extracted from the dataset which showed dependency and correlation between the AlcoholMortality, DiabetesDiagnosis, YLL_Alcohol, and CancerDiagnosis attributes. A strong rule was also generated between CancerDiagnosis and CurrentSmokers. However, having a lift value of 0.973, the rule had a negative correlation, which meant the increase in CancerDiagnosis is negatively correlated to Smoking. Based on the achieved results, we can say the application of exploratory unsupervised data mining techniques to the dataset yielded meaningful insights which satisfies the aim of the study.

Keywords: Unsupervised Learning, Association Rule Mining, Clustering, Discretization, Pattern Recognition.

I. INTRODUCTION

Public health can be described as the systematic process of preventing diseases, prolonging life, and preventing impairments through a collaborative effort by the society (Gray, Pilkington, Pencheon, & Jewell, 2006). It is described as a collective effort since everyone in the society is exposed to the same degree if factors that can affect the health or wellbeing of a member/individual is introduced to the society. Following the Health and Social Care Act of 2012, the responsibility for the administration and management of public health services in the United Kingdom (UK) was transferred back to the local governments and Public Health England on April 1, 2013, thereby making them democratically accountable stewards of their local populations' wellbeing. This was based on the reasoning that the local authorities were better positioned to oversee public health concerns in the society (Atkins, Kelly, Littleford, Leng, & Michie, 2019)and make policies that would improve the health of their local populations. The re-delegation of public health functions to the local government is one of the most significant extensions of local authorities' powers, as it provides a means for changing existing objectives from that of treating sickness as adopted by the NHS to proactively promoting health and wellbeing (Local Government Association, 2019).

II. PROBLEM STATEMENT

To aid the local authorities in understanding the public health circumstances of their populations, the Public Health England publishes reports which comprise information on a range of indicators that reflect the Health Profiles in each local authority in England. These health profile reports present a snapshot of indicators that summarise the health of the population in each local authority in England and the populations exposure level to harmful lifestyle practices. The reports are intended to help the local authorities and public health services in making proactive plans and policies aimed at improving the health of their local populations and reduce health inequalities. However, most local authorities have found it difficult to extract advanced analytic intelligence from these data sources due to a shortage of data analytics experts or possibly not knowing the potentials of such optimization. This study seeks to answer the question: how can data mining techniques can be used to optimise these health profile data for local authorities in England and extract advanced analytic insights that would reveal the relationship between lifestyle behaviours and diseases that affect the health of their populations?

Motivation: Health literacy involves people's comprehension, motive, and capacity to access, evaluate, and apply health-related information in making decisions centred on health care, disease prevention, and improvement of life quality (Brandt, et al., 2019). In the context of this research, health literacy encompasses the degree to which local authorities can obtain, process, and understand health-related information needed to design policies or take proactive measures to improve the health of their local populations and reduce health inequalities. In order to achieve this, there is a need to make evidence-based health information available to authorities charged with the administration and management of public health services. As a data scientist, the motivation for this work is the desire to bridge the gap created due to the shortage of data analytics experts in the local authorities by applying exploratory data mining techniques to these health profile reports to extract advanced analytic insights that would aid the understanding of possible relationships between lifestyle behaviour and diseases presented in the health profile reports. This would thus provide evidence-based health literacy for policymaking.

III. RELATED WORK

Farrell et al. in their work applied multivariate analysis using statistical packages for social sciences (SPSS) to investigate the association between variables comprising sociodemographic and health risk factor groups, and the dietary risk factors for overweight and obesity in Samoa women (Farrell, et al., 2019). Having used cross-sectional data, they couldn't establish a relationship between both sociodemographic and health behaviour factors, and diet. In this research, we will apply data mining techniques using cluster analysis and association rule mining to extract association rules between variables representing lifestyle behaviour and diseases.

Ana-Maria Talos (Talos, 2016) in her study was able to establish strong relationships between lifestyle and Ialomita county population health status by investigating the importance of lifestyle in influencing the population health status compared to other important determinants such as demographic and socio-economic indicators. The methodology involved processing data from spatial analysis and health survey using SPSS21, a statistical package.

(Brandt, et al., 2019) In his research examined the association of Health Literacy (HL) and the use of alcohol and tobacco among Austrian adolescents to provide evidence-based guidance for the development and improvement of primary prevention programs. All data analysis was conducted using the Mplus 7.3 software, which is a statistical modelling program that performs data analysis. The result shows that the more difficult it was for the respondents to find, understand, appraise, and apply health-related data, the more cigarettes they have smoked in their entire life. Results for alcohol consumption also showed that the lower the respondents HL, the more frequent was their alcohol consumption.

Li et al. investigated the association between sociodemographic, psychosocial, lifestyle factors, and self-reported health (SRH) data in migrant Chinese labourers (Li, et al., 2017). They applied multilevel logistic regression analyses to investigate the association between the variables. Although their research seeks association between the variables, it's context is quite different from this study which hopes to identify patterns suggestive of a correlation and dependence among variables.

Vinci et al. implemented a cluster analysis on sociodemographic and lifestyle factors among adults with excess weight (Vinci, et al., 2019). They investigated how overweight individuals differ from people with normal weight across sociodemographic and lifestyle indicators. The study also aimed to identify discrete clusters of overweight participants based on sociodemographic and lifestyle factors. In this research, the clustering techniques have been applied to the local authority health profiles to group the local authorities into clusters based on their similarities and dissimilarities in attributes, aimed at providing local authorities with advanced intelligence to aid policymaking that would address specific needs of each cluster. Chen et al. applied data mining techniques to available data on established risk factors for preterm birth in a bid to extract meaningful patterns that will explain the anomaly which is a major cause of perinatal morbidity and mortality (Chen, Chuang, Yang, & Wu, 2011). The Neural network is used to extract the top 15 factors linked to preterm birth. A decision tree C5.0 is then used to classify these factors based on weights to generate rules. This method is a supervised mining technique as the data has a class label that determines if the baby is born preterm. In our research, we have employed exploratory unsupervised mining techniques using clustering and association rule analysis to probe our data for interesting patterns and subsequent strong association rules.

Cuesta et al. illustrated the potentials of applying data mining techniques in understanding complex public health concerns. They performed decision tree analysis to see how individual and neighbourhood level factors such as body mass index (BMI), lot trash, social disorder, hours outside, psychological distress, physical activity variables, etcetera influence urban diabetes and asthma (Cuesta, Coffman, Branas, & Murphy, 2019). Being that decision trees are mainly used for predictive analysis, a better technique to identify these individual and neighbourhood health indicators associated with asthma and diabetes in the urban population would have been an association rule mining technique, which we would apply in this study.

Most of these studies have been conducted using statistical packages for social sciences (SPSS) and supervised mining techniques to analyse datasets with dependent variables and possibly class attributes. This research seeks to apply unsupervised data mining techniques to a health profile dataset with independent attributes to investigate the possibility of extracting advanced intelligence that will aid the understanding of relationships between lifestyle behaviours and associated diseases, and also aid health authorities in policymaking.

IV. METHODOLOGY

A. Data

Our data sets were obtained from Fingertips, a web platform owned by Public Health England (PHE) which provides access to a wide range of health-related data limited to England. The data on fingertips is open source and publicly available.

To obtain a representative dataset for this study, attributes were extracted from three datasets between 2011 - 2018:

a. The local authority health profile dataset provides an overview of health for each local authority in England. It contains data on a variety of indicators highlighting issues that affect health.
b. The local alcohol health profile data consists of indicators which provide insight into the health and social consequences of alcohol misuse.
c. The local tobacco control profile data consists of indicators which present information on the extent of tobacco use, tobacco-related harm, and measures being taken to reduce this harm at the local authority levels.
d. Data objects for 324 local authorities were extracted from the primary datasets and data pre-processing to integrate, reduce, and transform the data was carried out using Microsoft Transaction-SQL and R-Studio.

B. Unsupervised Data Mining

Data mining involves Knowledge Discovery from Databases (KDD), which seeks to uncover hidden patterns by evaluating relationships in the data objects. The technique employed in this study is called unsupervised because there is no pre-existing relationship to be learnt from. There are no output variables to be predicted as characterized with the supervised mining method as the goal of the process is to unveil hidden patterns in the data based on existing relationships between the data attributes. For this project, we would apply unsupervised techniques such as cluster analysis and association rule mining.

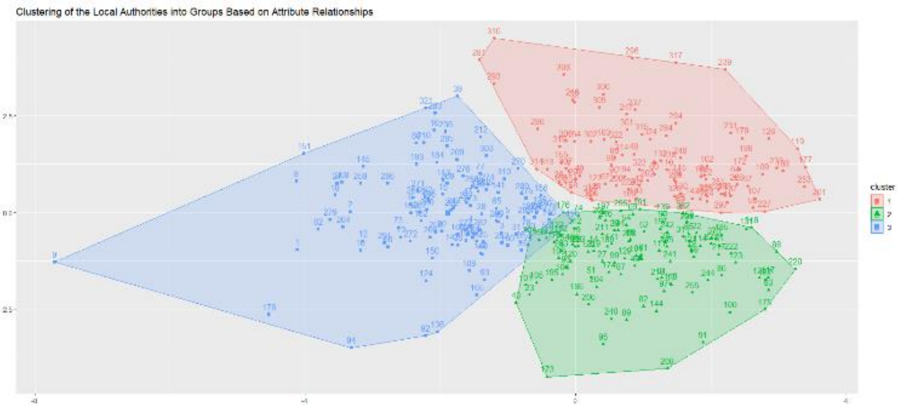
C. K-Means Clustering using the Euclidean Distance Measure.

In the clustering process, relationships between local authorities are formed based on similarities in attribute values, such that local authorities are grouped into clusters depending on similarities and dissimilarities in their values as compared to cluster centroids. This results in clusters where members of a cluster have high intra-cluster similarity and the different clusters have high inter-cluster dissimilarity. The motivation for this analysis is to see how clustering could be employed in assessing the exposure of local authorities to the social determinants of health and in policymaking/resource allocation based on regional disease impact assessments.

For a dataset to form meaningful clusters, it's data points should not be uniformly distributed. The get_clust_tendency () function from the factoextra package is used to assess the clustering tendency of the dataset. This function uses the Hopkins-Statistic to test the spatial randomness of an attribute as distributed in space, this is a spatial statistic measure for detecting cluster tendency in a dataset. A dataset contains meaningful clusters if its Hopkins-Stat is below 0.5 and tending to zero, which means the dataset is not uniformly distributed and thus the alternative hypothesis is adopted. The function also generates an ordered dissimilarity image (Han, Kamber, & Pei, 2011).

The clustering algorithms used for this study requires the number of clusters is specified. The fviz_nbclust () function from the factoextra package is therefore applied to the scaled dataset to determine the optimal number of clusters using the within-cluster sum of squares (WSS) method. This function iteratively creates clusters and plots the WSS of these clusters as a graph. Usually, the WSS or within-cluster variance drops as more clusters are formed. The technique to determine the right number of clusters is to select the cluster number at the point where there is a bend in the plot.

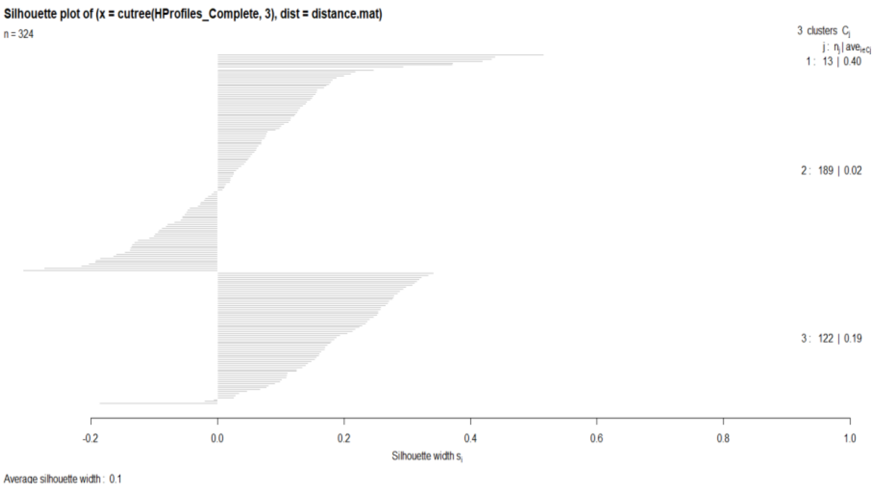
With a Hopkin-statistic value of 0.245, the alternative hypothesis is adopted; this means the dataset has meaningful clusters and is not uniformly distributed. The ordered dissimilarity image (ODI) shows areas of high dissimilarity colored blue while the regions colored red are areas of low dissimilarity and therefore high similarity. A close observation will reveal the areas in red looking as pockets which are clusters while the blue points seem to be boundaries and possible points of overlap between inter-cluster values.



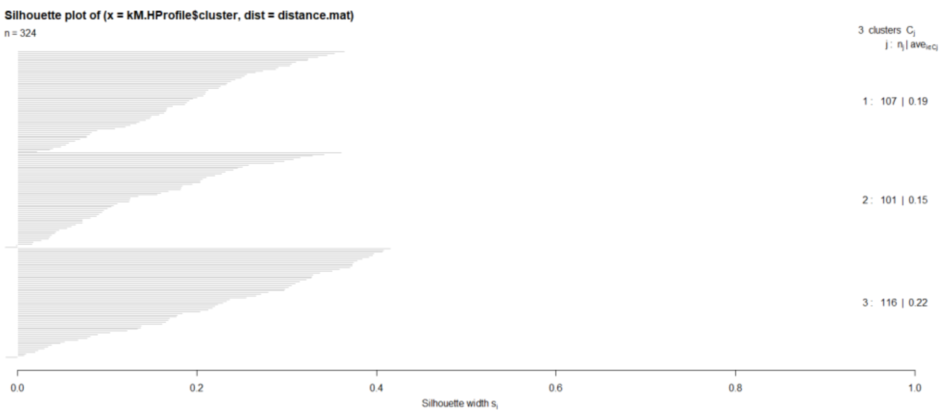
K-Means Clustering of the 2015 Health Profile Data The scree plot reveals that as the number of clusters increases, there is a drop in the within-cluster variance due to a higher similarity index between data objects. The plot shows that 3 clusters are the optimum number of clusters existing in the dataset based on a subjective inspection of the plot which reveals an elbow point at cluster 3 on the x-axis. From the 4th cluster, we can see the clusters having a somewhat equal WSS (that is, a higher similarity index).

To assess the quality of the clusters, the intrinsic method is employed. This method is used to assess cluster quality when prior knowledge of possible clusters (that is, the class attribute of the data set) is not known (Han, Kamber, & Pei, 2011)). This involves using the internal structural information of the clusters to evaluate the goodness of the method employed by evaluating how well the clusters are separated and the within-cluster density. In this study, the silhouette plot is used to validate the quality of clusters by calculating the silhouette width of each cluster. Result of the silhouette plot shows cluster 1 with 13 members as having the largest silhouette (Si) width of 0.40. This means the cluster is highly compact. Cluster 2 has the lowest Si width value of 0.02, and this means the data points within this cluster are not compactly packed. Examining cluster 2 in the Si plot reveals that a good number of its data objects have negative Si values which may signify outliers or wrongly classified data objects, meaning that the data points are closer to points in other clusters than it is to members of its cluster (Rousseeuw, 1987). Figure 4.5 below is a silhouette plot for the hierarchical clustering of our dataset.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com



Silhouette Plot of the Hierarchical clustering of the 2015 Health Profile Data. The silhouette plot for the k-means clustering was also obtained to validate the cluster quality. Results show a more distinct separation of the clusters with cluster 3 having the highest Si width values of 0.22. Outliers in cluster 2 are represented by the data points having negative silhouette values. These outliers may also be due to the overlap that occurs between data points in clusters 2 and 3 as seen in Figure 4.4 above. Figure 4.6 below gives a pictorial representation of the silhouette plot from the k-means clustering.



Silhouette Plot of the K-Means clustering of the 2015 Health Profile Data.

D. Association Rule Mining

Association rule mining is a technique built on probability principle that seeks to identify associations, patterns, or co-occurrence between attributes in a dataset. It uncovers recurring relationships between attributes in rows of data objects and based on data attributes forms rules which are called association rules. The aim here is to search for associations between the attributes in the health profile data and thus provide evidence-based rules that will lay the foundations for policymaking to protect and sensitise the population on the negative effects of tobacco and alcohol use.

The health profile data used in this study is made up of attributes such as cancer diagnosis, diabetes diagnosis, STI diagnosis, Smoking Prevalence, Alcohol Mortality, Years of life lost due to alcohol-related incidences, Current Smokers, Never Smoked, and Population distribution data. Cancer, diabetes, and sexually transmitted infections (STI) have been “associated” with lifestyle behaviours linked to tobacco and alcohol abuse. The motivation for this work stems from a desire to investigate this believed association by identifying strong association rules which show correlation supportive of the belief that these lifestyle behaviours cause these diseases. For this study, data mining techniques using the apriori algorithm from the arules package will be applied to our dataset to uncover frequent attribute-sets.

Exploratory analysis shows the dataset contains attributes with continuous values which cannot be mined for interesting patterns and rules. Therefore, pre-processing the data into a format suited for the mining algorithm to be employed is a crucial first step.

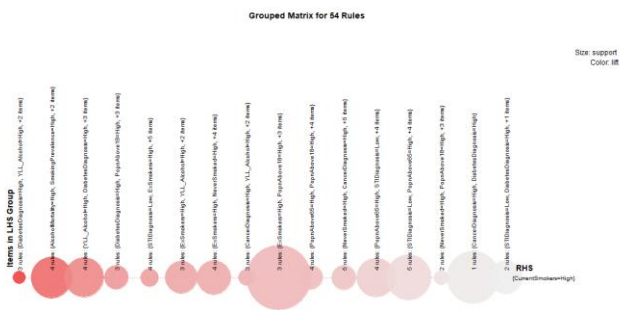
e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

To achieve our objectives, a discretization technique is applied to the dataset to group the numerical values into bins, intervals, or supervised clusters. In this study, the K-Means algorithm is used for our data discretization. Here the K-Means algorithm is applied manually to individual attribute columns instead of the whole dataset.

To discretize our data, we chose the labels/bins “High” and “Low”; “High” meaning the attribute value for the local authority is significant while the label “Low” means the attribute value for the local authority is non-significant.

In determining the number of discretized clusters to create, we put into consideration the fact that since all the local authorities in our dataset have values(a level of exposure), a measure to ascertain the significance level of the values is necessary. Using the k-means () function from the cluster package, each attribute column is discretized into three clusters: a, b, and c. Using the mean values of the clusters, two clusters with the largest mean values were classified significant and thus labelled “High”, while the cluster with the lowest mean was classified non-significant and labelled “Low”. The cbind () function is then used to bind the vectors containing the attribute cluster assignment of each local authority into a data frame column-wise. Next the data frame with values occurring as integers (cluster numbers) is converted to character values using the as.character () function. The next step would involve the application of the apriori () function to the dataset.

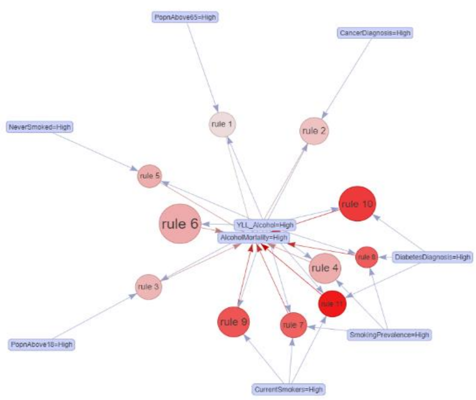
To mine the dataset for frequent attribute-sets, the apriori () function from the arules package (Hahsler, Buchta, Hornik, & Chelluboina, 2011) is applied to the dataset using the default parameter setting of support = 0.1, confidence = 0.8, and maximum rule length of 10. Applying the summary function to the above result shows a total of 4,397 rules are generated.



Grouped Matrix for 54 association rules at min.support = 0.3, Min Confidence = 0.7, and RHS attribute = CurrentSmokers (High)

A brief inspection of these rules shows a lot of uninterestingness in the generated rules. To obtain stronger association rules, the minimum support and confidence values are tweaked using the parameter () function within the apriori () function, such that rules which satisfy the user-defined minimum support and confidence values are generated.

To make the rule generation more interactive, the ruleExplorer () function from the arulesViz package (Hahsler, arulesViz: Interactive Visualization of Association Rules with R, 2017) is used. Using the ruleExplorer () function, data attributes on the right-hand-side (RHS) of the rule can also be interchanged to measure dependencies.



e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

Graph of association rules at Min. Support = 0.35, Min. Confidence = 0.7, and RHS attribute = AlcoholMortality

V. RESULTS AND DISCUSSION

E. Cluster Analysis

The K-Means clustering of the data using the optimum cluster value reveals the distribution of the local authorities into 3 clusters. Results show that the first, second, and third clusters contain 107, 101, and 116 local authorities respectively. Looking at the results of our experiments, we can conclude that the K-Means method performs adequately in clustering our dataset. Based on the cluster means for the different clusters across the attributes, we can determine the attributes that played a major role in cluster formation and make the following statements:

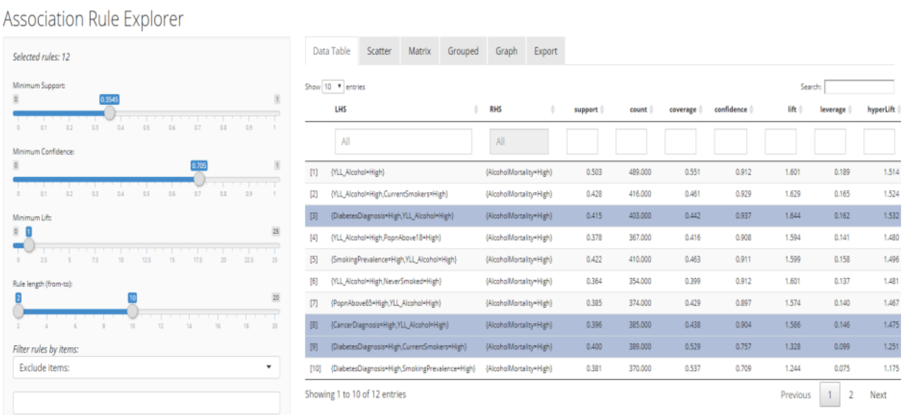
i. Local authorities in cluster 3 have High SmokingPrevalence and CurrentSmokers compared to cluster 2 and 1.
ii. Local authorities in cluster 3 have a larger population of people diagnosed with diabetes compared to clusters 1 and 2.
iii. Local authorities in cluster 3 have a high rate of alcohol-related mortality(AlcoholMortality) and years of life lost due to alcohol-related incidences(YLL_Alcohol) compared to clusters 1 and 2.

Based on these statements, we can therefore conclude that by unveiling the relationships between local authorities, cluster analysis as an exploratory data mining technique provides advanced analytic insights which can be used in policymaking and allocating resources to the sensitization of the local authority populations in cluster 3 on the harmful effects of tobacco and alcohol, providing services to improve the life of people with diabetes, and making more stringent policies on the use of tobacco in public especially in local authorities with high cancer diagnosis.

F. Association Rule Mining

The attribute-sets shows that the generated rules are strong association rules as they are frequent attribute-sets and satisfy the minimum support and confidence values we have set. The rules also show dependency and correlation among the attributes with lift values between the range of 1.290 to 1.025. This means that Populations with a High number of Current Smokers are likely to have a High number of CancerDiagnosis, Diabetes Diagnosis, and Alcohol-related mortalities. We also see high cancer and diabetes diagnosis occurring in current smokers in populations above 65.

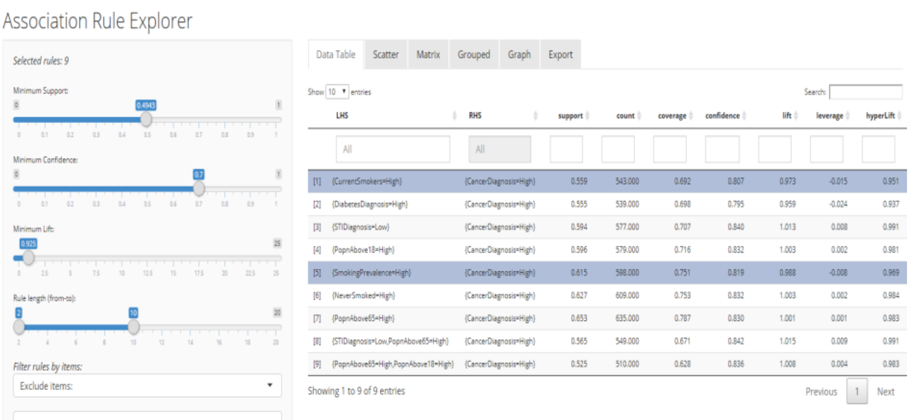
Further analysis to investigate the relationship between alcohol and diabetes shows strong association rules between diabetes diagnosis, years of life lost due to alcohol (YLL_Alcohol) and Alcohol-related mortality. The lift values of 1.644 and 1.586 indicate dependency and correlation between the attributes since the rules show a high record of alcohol-related mortality diagnosis in areas with high diabetes and high cancer diagnosis. This suggests a plausible and significant association indicating that alcohol consumption can lead to diabetes and Cancer.



	LHS	RHS	support	count	coverage	confidence	lift	leverage	hyperLift
[1]	{YLL_Alcohol=High}	{AlcoholMortality=High}	0.503	489.000	0.551	0.912	1.601	0.189	1.514
[2]	{YLL_Alcohol=High,CurrentSmokers=High}	{AlcoholMortality=High}	0.428	416.000	0.461	0.929	1.629	0.165	1.524
[3]	{DiabetesDiagnosis=High,YLL_Alcohol=High}	{AlcoholMortality=High}	0.415	403.000	0.442	0.937	1.644	0.162	1.532
[4]	{YLL_Alcohol=High,PopnAbove18=High}	{AlcoholMortality=High}	0.378	367.000	0.416	0.908	1.594	0.141	1.480
[5]	{SmokingPrevalence=High,YLL_Alcohol=High}	{AlcoholMortality=High}	0.422	410.000	0.463	0.911	1.599	0.158	1.496
[6]	{YLL_Alcohol=High,NeverSmoked=High}	{AlcoholMortality=High}	0.364	354.000	0.399	0.912	1.601	0.137	1.481
[7]	{PopnAbove65=High,YLL_Alcohol=High}	{AlcoholMortality=High}	0.385	374.000	0.429	0.897	1.574	0.140	1.467
[8]	{CancerDiagnosis=High,YLL_Alcohol=High}	{AlcoholMortality=High}	0.396	385.000	0.438	0.904	1.586	0.146	1.475
[9]	{DiabetesDiagnosis=High,CurrentSmokers=High}	{AlcoholMortality=High}	0.400	389.000	0.529	0.757	1.328	0.099	1.251
[10]	{DiabetesDiagnosis=High,SmokingPrevalence=High}	{AlcoholMortality=High}	0.381	370.000	0.537	0.709	1.244	0.075	1.175

Association rules at Min. Support = 0.35, Min. Confidence = 0.7, and RHS attribute = AlcoholMortality.

Among our finding are a set of two interesting itemset rules which do not seem to support the belief that smoking is correlated to cancer:



	LHS	RHS	support	count	coverage	confidence	lift	leverage	hyperLift
[1]	{CurrentSmokers=High}	{CancerDiagnosis=High}	0.559	543.000	0.692	0.807	0.973	-0.015	0.951
[2]	{DiabetesDiagnosis=High}	{CancerDiagnosis=High}	0.555	539.000	0.698	0.795	0.959	-0.024	0.937
[3]	{STIDiagnosis=Low}	{CancerDiagnosis=High}	0.594	577.000	0.707	0.840	1.013	0.008	0.991
[4]	{PopnAbove18=High}	{CancerDiagnosis=High}	0.596	579.000	0.716	0.832	1.003	0.002	0.981
[5]	{SmokingPrevalence=High}	{CancerDiagnosis=High}	0.615	598.000	0.751	0.819	0.988	-0.008	0.969
[6]	{NeverSmoked=High}	{CancerDiagnosis=High}	0.627	609.000	0.753	0.832	1.003	0.002	0.984
[7]	{PopnAbove65=High}	{CancerDiagnosis=High}	0.653	635.000	0.787	0.830	1.001	0.001	0.983
[8]	{STIDiagnosis=Low,PopnAbove65=High}	{CancerDiagnosis=High}	0.565	549.000	0.671	0.842	1.015	0.009	0.991
[9]	{PopnAbove65=High,PopnAbove18=High}	{CancerDiagnosis=High}	0.525	510.000	0.628	0.836	1.008	0.004	0.983

Association rules at Min. Support = 0.50, Min. Confidence = 0.7, and RHS attribute = CancerDiagnosis.

CurrentSmokers ➔CancerDiagnosis

SmokingPrevalence ➔ CancerDiagnosis

Although the association rules are strong (having surpassed the set minimum support and confidence values of 0.5 and 0.7), their lift values of 0.973 and 0.988 respectively are below 1. This means that there is a negative correlation between cancer diagnosis, current smokers, and high smoking prevalence.

VI. CONCLUSION

This aim of this research was to extract advanced intelligence which can be used to formulate health policies by providing an understanding of the relationship between lifestyle behaviours and diseases affecting the health of the population. Based on our application of exploratory analysis on the health profile data using data mining tools, the knowledge discovery process uncovered meaningful patterns which showed strong relationships between lifestyle behaviours and diseases affecting the health of the population. Results from our clustering analysis successfully grouped local authorities into clusters defined by their exposure levels to the investigated diseases and lifestyle behaviours. Looking at the results, health authorities can assess the impact of existing policies, assess factors which might have made the policies fail in some local authorities, and also prepare budgets to cater for local authorities based on their needs as evidenced by the cluster they belong. The strong association rule mining generated interesting rules which can be used to effect positive changes in the health of populations within England. Alcohol was identified as a major cause of diabetes and cancer based on the dependencies and positive correlation discovered in the association rules. This is an interesting rule which can be applied in campaigns against alcohol abuse. The trend analysis on tobacco profiles confirms the successful effects of existing policies and sensitization on tobacco use. Although the negative correlation between cancer and tobacco does not support the general belief that smoking causing most cancers, this study provides evidence-based facts which prove that association rule generalisation is not ideal as there may be secondary factors that make certain rules unique or adapted to a particular area. This research has therefore shown that using exploratory data mining techniques, advanced analytic insights can be extracted from population health profile data which would provide local authorities evidence-based facts to aid the development of health policies and sensitization campaigns that would help improve the lives of their population.

VII. REFERENCES

[1] Atkins, L., Kelly, M. P., Littleford, C., Leng, G., & Michie, S. (2019). From the National Health Service to Local Government: Perceptions of Public Health Transition in England. Public Health, 174, 11-17.

[2] Brandt, L., Schultes, M. -T., Yanagida, T., Maier, G., Kollmayer, M., & Spiel, C. (2019, May). Differential associations of health literacy with Austrian adolescents' tobacco and alcohol use. Public Health, 174, 74-82.

[3] Chen, H.-Y., Chuang, C.-H., Yang, Y.-J., & Wu, T.-P. (2011). Exploring the risk factors of preterm babies using data mining. Expert Systems with Applications, 5384-5387.

[4] Cuesta, H. A., Coffman, D. L., Branas, C., & Murphy, H. M. (2019). Using decision trees to understand the influence of individual- and neighbourhood-level factors on urban diabetes and asthma. Health & Place, 58, 109-119.

[5] D'monte, S., & Panchal, D. (2015, July o6). Data mining approach for diagnose of anxiety disorder. International Conference on Computing, Communication & Automation, 124-127. doi:doi: 10.1109/CCAA.2015.7148357

[6] Farrell, P., Negin, J., Awoke, M., Thow, A. M., Taua, M., Faumuina, T., . . . Richards, J. (2019). Associations between sociodemographic and behaviour factors, and dietary risk factors for overweight and obesity, in Samoan women. Appetite, 134, 155-161.

[7] Gray, S., Pilkington, P., Pencheon, D., & Jewell, T. (2006, March). Public Health in the UK:Success or Failure? Journal of the Royal Society of Medicine, 99(3), 107-11.

[8] Hahsler, M. (2017). arulesViz: Interactive Visualization of Association Rules with R. R Journal, 9(2), 163-175. Retrieved from https://journal.r-project.org/archive/2017/RJ-2017-047/RJ-2017-047.pdf>

[9] Hahsler, M., Buchta, C., Hornik, K., & Chelluboina, S. (2011). The arules R-Package Ecosystem:Analyzing Interesting Patterns from Large Transaction Datasets. Journal of Machine Learning Research, 12, 1977-1981. Retrieved from http://jmlr.csail.mit.edu/papers/v12/hahsler11a.html

[10] Han, J., Kamber, M., & Pei, J. (2011). DATA MINING Concepts and Techniques (3 ed.). Amsterdam: Morgan Kaufmann.

[11] Li, C.-C., Meng, X.-H., Wang, J.-R., Ma, H.-J., Chen, C., & Liu, Y.-Q. (2017). Association between socidemographic, psychosocial, lifestyle factors, and self-reported health among migrant laborers in China. Journal of the Chinese Medical Association, 204-211.

[12] Local Government Association. (2019). Public health. Retrieved 08 20, 2019, from Local.gov.uk: https://www.local.gov.uk/topics/social-care-health-and-integration/public-health

[13] Ordonez, C. (2006, November 11). Comparing association rules and decision trees for disease prediction. HIKM'06 Proceedings of the international workshop on healthcare, information, and knowledge management, 17-24. doi:10.1145/1183568.1183573

[14] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpetation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

[15] Talos, A.-M. (2016). Influence of population lifestyle on local health profile. Case study: Ialomita county. Internatin Conference - Environment at a Crossroad: SMART approaches for a sustainable future. 32, pp. 311-317. Elsevier.

[16] Vinci, L., Krieger, J.-P., Braun, J., Pestoni, G., Bender, N., Rohrmann, S., . . . Staub, K. (2019). Clustering of Sociodemographic and Lifestyle Factors among Adults with Excess Weight in Multilingual Country. Nutrition, 177-185.