Research Article

BIOMEDICAL Journal of Scientific & Technical Research

ISSN: 2574 -1241

DOI: 10.26717/BJSTR.2023.48.007641

Profiling Obese Subgroups in National Health and Nutritional Status Survey Data using Machine Learning Techniques – A Case Study from Brunei Darussalam

Usman Khalil[1]*, Owais Ahmed Malik[1,2], Daphne Teck Ching Lai[1] and Ong Sok King[3]

[1]*School of Digital Science, Universiti Brunei Darussalam, Jalan Tungku Link, Brunei*

[2]*Institute of Applied Data Analytics, Universiti Brunei Darussalam, Jalan Tungku Link, Brunei*

[3]*Public Health Services, Ministry of Health, Brunei Darussalam and PAPRSB Institute of Health Sciences, Universiti Brunei Darussalam, Jalan Tungku Link, Brunei*

***Corresponding author:** Usman Khalil, School of Digital Science, Universiti Brunei Darussalam, Jalan Tungku Link, Gadong BE1410, Brunei Darussalam

ARTICLE INFO

Received: January 20, 2023

Published: February 01, 2023

Citation: Usman Khalil, Owais Ahmed Malik, Daphne Teck Ching Lai and Ong Sok King. Profiling Obese Subgroups in National Health and Nutritional Status Survey Data using Machine Learning Techniques – A Case Study from Brunei Darussalam. Biomed J Sci & Tech Res 48(3)-2023. BJSTR. MS.ID.007641.

ABSTRACT

National Health and Nutritional Status Survey (NHANSS) is conducted annually by the Ministry of Health in Negara Brunei Darussalam to assess the population's health and nutritional patterns and characteristics. The main aim of this study was to discover meaningful patterns (groups) from the obese sample of NHANSS data by applying the data reduction and interpretation techniques. The mixed nature of the variables (qualitative and quantitative) in the data set added novelty to the study. Accordingly, the Categorical Principal Component (CATPCA) technique was chosen to interpret the meaningful results. The relationships between obesity and lifestyle factors like demography, Socio-Economic status, physical activity, dietary behavior, history of blood pressure, diabetes, etc., were determined based on the principal components generated by CATPCA. The results were validated with the help of the split method technique to counter-verify the authenticity of the generated groups. Based on the analysis and results, two subgroups were found in the data set, and the salient features of these subgroups have been reported. These results can be proposed for the betterment of the health care industry.

Keywords: NHANSS; Data Mining; Machine Learning; Categorical Principal Component Analysis; CATPCA; NCD; Obesity

Introduction

Obesity is one of the non-communicable diseases that is a condition of being overweight or a major nutritional disorder that has become a worldwide epidemic. Its growth has been projected at 40% in the upcoming decade [1]. It is often defined simply as a condition of abnormal or excessive fat accumulation in adipose tissue to the extent that health may be impaired [2]. Not only is being obese problematic but with that comes the risk and complications of other non-communicable diseases (NCDs) that can be life-threatening if not taken care of at the right time, e.g., hypertension/high blood pressure, diabetes, etc. [3]. Diseases can be communicable and non-communicable diseases; the names refer to the type as the diseases spread (by different means) from one individual to another, such as pneumonia, malaria, hepatitis-A and C, HIV/AIDS, measles, etc.

At the same time, the latter is not transmissible directly from one person to another. Obesity, cancer, heart disease, diabetes mellitus, cerebrovascular disease, hypertension, high blood pressure, high cholesterol levels, etc., are all NCDs [1,2].

WHO Obesity Classification

WHO defined obesity as an accumulation of excessive body fats in tissues to the extent that health may be impaired. BMI measures it in kg/m2 [2,4,5]. It further defined the Overweight and obesity for adults as follows; BMI more than equal to 25 kg/m2 for Overweight; and BMI more than equal to 30 kg/m2 for Obese [2,6-8]. Obesity has been further classified as BMI more than equal to 30kg/m2. It includes an additional sub-division as BMI more than equal to 30kg/m2 and less than equal to 34.9 kg/m2 for obese class-I, BMI more than equal to 35kg/m2 and less than equal to 39.9 kg/m2 for obese class-II and BMI more than equal 40 kg/m2 for obese class-III [2-4,9-11]. However, this classification does not completely consider the population-level heterogeneity and cannot identify the variations among obese individuals. There is evidence of the association of obesity with other factors, including demographics, nutritional habits, and individuals' physical activity [7,8]. In our case, Body Mass Index (BMI) was calculated and inserted into the dataset as a variable feature to study the characteristics of obese people and the prevalence of obesity [1,6]. This survey also used the same variable features like demographic status, diet patterns, and physical activity together with the history of raised blood pressure, diabetes, and raised cholesterol with BMI measurements as done in the studies carried out in the past [8].

ASEAN Strategic Framework on NCDs

Taking a step ahead, the Brunei Darussalam government has taken significant measures to handle the NCD-related issues in its population [3]. Following the World Health Organization (WHO) Global Action Plan to control the prevalence of non- communicable diseases (NCDs) and the ASEAN Strategic Framework on Health Development [10]. The government has well anticipated the execution of the plan and has initiated a Multisectoral Action Plan on NCD (BruMap-NCD) 2013-2018 to control NCDs and related risk factors. It includes a ban on all kinds of smoking products in the country with a 30% reduction in smoking prevalence and a 10% reduction in physical inactivity prevalence by 2018 from the 2013 level [3]. According to this National Action Plan on the Prevention and Control of Noncommunicable Diseases (BruMAP-NCD) 2013-2018, Brunei's 1st National Nutritional Status Survey (NNSS) was carried out in the year 1997. Around 32% of the population was overweight, and 12% were obese among 20 years old and above [2,6]. This obese percentage, in particular, was increased by more than double to 27.2% in the year 2011 [3,11,12]. The current statistics show that around 61% of Bruneians are overweight and obese, the highest rate in ASEAN [2,5,6,12].

Obesity Prevalence in Brunei Darussalam

Brunei Darussalam is an oil & gas producing country and is one of the member countries in the ASEAN (Association of Southeast Asian Nations) organization. It is situated in Southeast Asia on the northern coast of Borneo Island, neighboring its borders directly with Malaysia. Its population is estimated at 417,200, with gross domestic products (GDP) per capita of USD 28,986 [6]. There has been a noticeable rise in non-communicable diseases (NCDs) as aforementioned [1], while obesity, one of NCDs, has been of major concern for its occurrence. The government has been targeting management and prevention from the grassroots level to overcome this problem, including childhood obesity. It requires long-term strategies and treating childhood obesity may likely help manage obese adults in the future [2,5]. The study's focus has been on the prevalence of obesity and the lifestyle factors affecting it. Past studies have shown that it has been one of the major risk factors causing other non-communicable diseases such as diabetes and cardiovascular problems [7-9]. The threat of obesity related NCDs, especially chronic kidney diseases (CKD), is preventable by educating the population about the risks of being obese and prevention through a healthy lifestyle [5]. In 2014, over 600 million adults aged 18 years and above were obese worldwide [5].

National Health and Nutritional Status Survey Data

The data was provided by the ministry of health Brunei Darussalam which runs parallel with the Brunei Darussalam Household Expenditure Survey (HES) 2010/2011 implemented by the Department of Economic Planning and Development, Prime Minister's Office [11]. NHANSS, the acronym for National Health and Nutritional Status Survey, is conducted annually to access the population's health and nutritional patterns and characteristics [11]. The data includes all the lifestyle aspects regarding demographics, Socio-Economic status, physical activity, and laboratory examinations. The Ministry of Health (MOH) designed the data collection process and carried it out in three phases. Sampling procedure, questionnaire development, database development, and testing. Like others, NHANSS is also a cross-sectional survey aimed at the population aged from 5-to 75 years old, with an initial target of 2184 participants from all the districts in Brunei Darussalam. All the health offices under the ministry of health were included for data collection, including Tertiary Care Hospitals, Health Offices in Districts, Health Clinics, and the Community Nutrition Centre were used as survey sites. Face-to-face interviews with parents and/or caregivers (for children) and participants themselves were conducted by trained dietitians/nutritionists and research assistants using a questionnaire booklet [11]. The measurements, including anthropometric indices such as weight, height, and waist circumference, were taken. Blood pressure readings were also noted for all respondents using standard methodology [13], while individuals aged 20 years and above were

additionally asked for biochemical measurements. Before the final data collection, a test run was carried out on the survey procedures and questionnaire to have standardized data collection [3,11,13,14].

Categorical Principal Component

The field of study interested in developing computer algorithms to transform data into intelligent action is known as machine learning [15]. Machine learning techniques have been used to explore the details mentioned above, which have been of great importance to extract the useful knowledge from the data that normally is received from a group of individuals through a survey or a questionnaire, or other health-related data collection techniques [15,16]. Categorical Principal Component Analysis (CATPCA) is one of the techniques applied to the data sets with more variables to reduce the dimensionality of the data set by ensuring as much variation as possible and, most importantly, applied to the set of qualitative and quantitative variables. The goal of the technique is to reduce an original data set into a smaller set of uncorrelated components (variables) that represent most of the information found in it. It removes a large number of correlated variables that may affect the interpretation of the patterns projected by the reduced variables. By dimensionality reduction, a few components with high variance interpret the patterns rather than many components with no or low variance. The choice of the technique was evident and applied to the NHANSS data set as discussed in Section 1.4 to generate meaningful patterns as far as obesity prevalence in the community was concerned. The review of the obtained results may help the health care industry to classify the characteristics of patients for a particular disease and to use that information to improve the protocols and procedures for the better treatment of patients by the clinicians and, most importantly, for the betterment of humanity in general [8,16,17].

Research Objectives

Focusing on the same idea, the objectives in this research were set as follows

- We explored the NHANSS data set and identified subgroups within the obese sample by implementing the machine learning technique.
- We resolve the pre-processing data issues by applying missing values analysis, imputation analysis, and data normalization techniques.
- We review and analyze the generated patterns by dimensionality reduction for obesity and the factors affecting it.
- We interpret and profile the salient characteristics of sub-groups based on result validation.
- Finally, we provide insightful reviews and discussions on generating potential recommendations and relevant information about the affecting factors of obesity to clinicians for preventive measures.

Paper Organization

The rest of the paper is organized as follows. Section 2 elaborates on the overall NHANSS obese sample, data processing issues, interpretation, analysis, and results from validation classification methods. Section 3 presents the CATPCA analysis and the validation process and finalizes the profiling to generate salient characteristics of the obese sample. Finally, a concise conclusion is presented in Section 0 at the end.

Methodology

The overview of the model methodology to carry out the study has been provided in Figure 1. Data was taken from the NHANSS – 2017 provided by the Ministry of Health, Negara Brunei Darussalam, representing data collection and selecting the variables in the first step. The second step follows the data pre-processing for any missing values or normalization issues so that data can be applied with the machine learning techniques. The categorical principal component analysis (CATPCA) extracted the components by reducing the dimensions and classifying the data. In this research, the classification technique was tested, and the process of validation was carried out to check the authenticity of the generated results. At the same time, the last step concludes the interpretation by profiling the observed classes. The results, validation, and profiling steps were carried out to understand and present the intelligible data for reporting. Since the motive was to find the meaningful patterns, Figure 1 shows the steps performed for identifying the subgroups of the obese in a given sample. The steps below mentioned were followed,

1) NHANSS ~ Obese Sample
2) Data Pre-Processing
3) Classification Method
4) Interpretation & Analysis (Results & Discussion)
5) Results Validation.

NHANSS ~ Obese Sample

As discussed in Section 1.4 and to study the characteristics of the obese population within the obese classes (I-II-III), the NHANSS data set (National Health and Nutritional Status Survey) was filtered with the number of people having BMI ≥ 30 kg/m2. Out of the total sample of 2184 records, 449 were filtered with 20.55% percent, and the required set of variables was chosen. A subset data set was chosen from the NHANSS data, whereas all the variables were included based on evidence-based research on obesity [7,13,18]. Since the obese sample had mixed variable types, the data type measurement for the variables was defined as quantitative and qualitative. It also added to

the study's novelty as not many studies on the obesity affecting factors have been carried out in the past with mixed variables data types. The level of measurement for quantitative variables was numeric, while for qualitative variables, the level of measurement was set either nominal (for not ordered data) or ordinal (for ordered data). In Step 3 in Figure 1, the machine learning technique was applied once the data was pre-processed. The CATPCA (categorical principal component analysis) was chosen for this study because of its ability to handle qualitative and quantitative data.

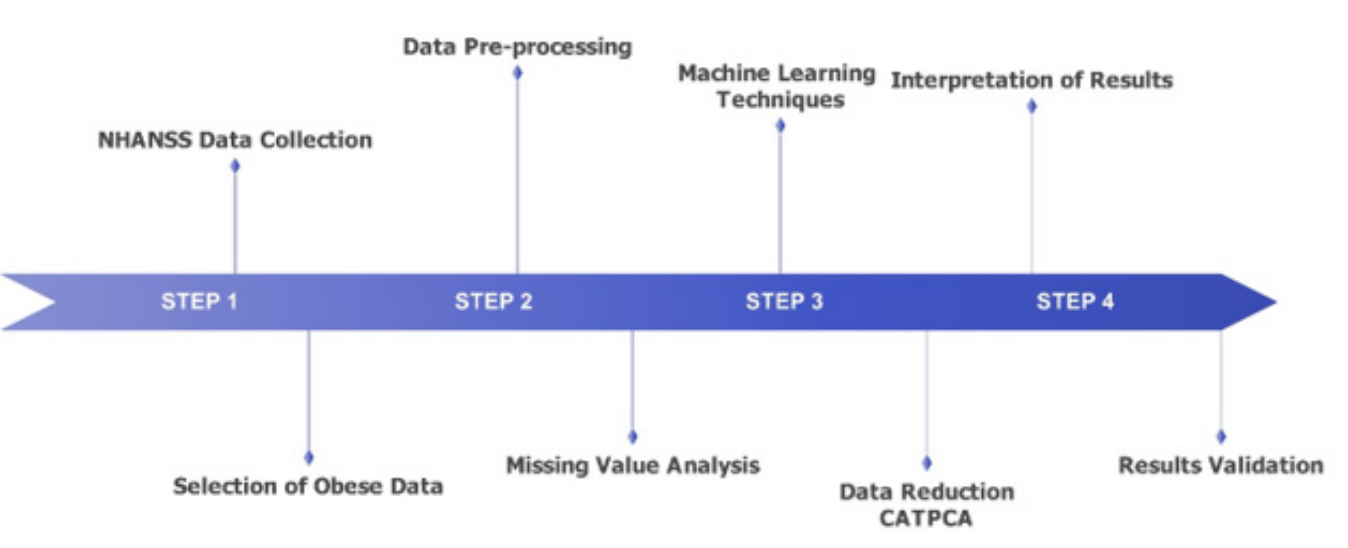

Figure 1: Classification Model Methodology.

Data Pre-Processing

Like the other surveys, the NHANSS is a cross- sectional survey conducted among all age groups in all four districts. Figure 2 lists the details for data collection [7,10,13]. As represented, 67.70% of the data was collected from Brunei Muara, the most densely populated, 17% from Kuala Belait, 12% from Tutong, and 3.30% from Temburong, being the lowest among all. A comprehensive questionnaire was prepared to note down the critical information, which was taken in several groups, such as;

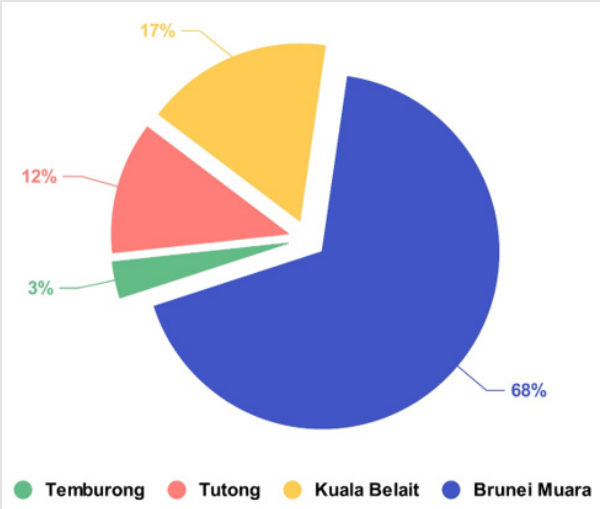

Figure 2: Data Collection across the Districts of Brunei Darussalam.

1) Demographics
2) Socio-Economic status
3) Medical / Smoking Status
4) Physical Activity Patterns
5) Anthropometric Measurements
6) Multiple Dietary Patterns
7) Bio-Chemical measurements on Adults and Children.

The NHANSS data set with 2184 instances and 88 variables were pre-processed for missing values after missing value analysis. Since the data set was already inputted with the missing values, the data set was further analyzed. This sample was representative of obese individuals from all three classes of obesity. It had 449 instances with 88 variables (86 excluding BMI and Obesity factor as evaluating factors) for the CATPCA analysis. The level of measurement for all the variables was ordinal, while there were 14 numeric (Scale) variables whose normalization was taken care of by SPSS with a normal distribution. The data points were processed in SPSS Ver. 20 and since the obese NHANSS data set was used with 86 variables and 449 instances, the representation of number of variables (m) were X1, X2, X3…….X86 i.e., m = 86 (e.g. X1= age years, X2 = Urban, X3 = DistCd and so on to…….. X86 = Salt93) while (n) represents the number of instances i.e., n=449 for obese sample. Demographic variables were selected, such as age, sex, marital status, etc., [19] for physical activity, the recreational activities such as vigorous or moderate activities were selected. For sedentary characteristics, the time spent watching TV and resting/reclining variables were taken as per their importance in the earlier studies [7,20]. Age was reported as a continuous variable (quantitative), while sex, ethnicity, etc., were reported as categorical variables (qualitative) [16]. Time spent for vigorous/moderate activities, watching TV, and resting/reclining time were taken as continuous variables since they indicate the importance of sedentary characteristics with time spent on it [19]. Dietary intake was self-reported through a questionnaire provided to the subjects in NHANSS data collection [11]. It was reported in categorical variables format in levels from 1 to maximum 7, varying for different variables with numbers 666 for not known and 999 for not applicable, respectively [19].

Results Validation

For validation purposes, the split method was used. The obese data set was divided into two data sets more cases from class I (187), with a percentage of 59.56%, is the highest among the classes, followed by class II (85) with a percentage of 27.07% being second highest and then class III (42) with a percentage of 13.37% being the lowest respectively.

Categorical Principal Component Analysis

Categorical principal component analysis (CATPCA) is applied to the data sets with more variables of mixed data types, i.e., qualitative and quantitative variables [21]. It reduces the dimensions of the data set by increasing variation as much as possible [18]. It is also referred to as Nonlinear Principal Component Analysis (PCA) [19], which works opposite of how PCA works. Nonlinear PCA reduces the observed variables to several uncorrelated variables [21]. If the measurement level of the variables is scaled to numeric, then PCA will be an alternative to CATPCA. Therefore, it would not be wrong to say that CATPCA is an alternate analysis technique to PCA when the analysis required is to find the patterns of variations in a single data set of mixed data types [22]. When PCA handles mixed quantitative and qualitative data, the qualitative data must be quantified and is known as nonlinear PCA [18]. The CATPCA solution maximizes correlations of the object scores with each of the quantified variables for the number of components (dimensions) specified. The CATPCA application is only available in IBM® SPSS®. If applied to all variables that are declared multiple nominals, CATPCA produces an analysis equivalent to a named train data set and test data set by a ratio of 70:30, which means 449 instances were divided by a ratio of 314:135, respectively. First, the results were generated by applying CATPCA on train data set with 314 instances. These results were compared to validate principal components by applying the same technique to the test data set (135 instances) later on. The descriptive statistics of obesity factors are presented in Table 1 with classes I, II, and III mentioned against obesity factors 1, 2, and 3 (1st column), respectively. The obesity factor was the representation of obesity classes in the data set. It can be seen that there were multiple correspondence analysis (MCA) run on the same variables, so CATPCA can be seen as a type of an MCA in which some of the variables are declared ordinal or nominal [22].

Table 1: Obese Sample ~ Train Data Set.

Obesity Factor	Obesity Class	Frequency	Percent	
Valid	1a	Obese - Class I	187	59.56%
	2	Obese - Class II	85	27.07%
	3	Obese -Class III	42	13.37%
	Total	314	100%	

Note: a. Mode

Component Extraction Methods: As discussed in the section above, one of the most important purposes of PCA / CATPCA methods is dimension reduction. In order to achieve the purpose, some criterion has to be applied, whose method may follow the same principles to reduce the dimensions. Selecting only a few Principal Components (PCs) that share less of the variance may not help as this might result in selecting too few PCs and reducing the dimensions a lot. Similarly, selecting all the PCs will also be of no use just because they explain most of the variance of the data and may not help as this might result in selecting most or all the PCs and not reducing the dimensions at all. It may not fulfill the essence of the dimension reduction method.

Component Extraction Criterions: The principal components that share the maximum variance should be the benchmark to select and reduce the dimensions. However, other defined criteria can be applied by looking at the data's nature. The different criteria available can be applied according to the nature of the data in the view. Four types of criteria can be used and are discussed below mentioned.

Eigen Value Criterion:

1) The proportion of Variance Explained Criterion
2) Minimum Communality Criterion
3) Scree Plot Criterion

Eigen Value Criterion: As per the eigenvalue criterion, a principal component must explain "one variable's worth," which would mean that the PCs must have an eigenvalue of 1 at least. Eigenvalue Criterion may be best suited for data sets with more than 20 and less than 50 variables if the data set has less than 20 variables. The criterion may choose too few principal components, and if the data set has more than 50 variables, then the criterion may choose too many principal components. In either case, it may not be feasible to analyze and later outline the characteristics of those chosen dimensions/components [18]. For instance, if there are the principal components PC1, PC2, and PC3 have eigenvalues $\lambda 1 = 1$, $\lambda 2 = 0.85$, $\lambda 3 = 0.075$ respectively then according to this criteria PC1 may be the only component retained, and the rest may be discarded. PC2 can also be retained as the eigenvalue is close to the threshold eigenvalue of 1, so in this case, two principal components may be retained, i.e., PC1 and PC2.

The Proportion of Variance Explained Criterion: This criterion mostly depends on the analyst who specifies the total number of principal components considering the variability. The PCs must be selected until the desired proportion of the variability explained is attained. The total proportion of the variability can be explained by Equation (2.1) below,

Where,

$$Z = \frac{\lambda_i}{m}$$

1) Z depicts the total variability in it,
2) λi depicts the eigenvalue for ith principal component.
3) m depicts the number of principal components.

The equation represents the proportion of variability in Z, which is explained by the ratio of ith eigenvalue for the ith principal component to the number of variables. For instance, if a data set has ten variables applied with CATPCA results with eigenvalues against respective principal components and the first principal component has an eigenvalue of $\lambda 1 = 4.901$; then, as per equation 2.1, since there are ten variables (m), the first component may explain 4.901/10 = 49.01% of the shared variance among the predictor variables. Suppose the required percent of shared variance among the predictor variables is 85%. In that case, more principal components may be added so that the desired number of components should attain the desired percent explained by the variability.

Minimum Communality Criterion: PCA / CATPCA does not present all the variance from the variables but only a proportion of the variance shared by the predictor variables. Communality plays an important role in extracting the proportion of a particular variable. Communality shows how beneficial the variable is for contributing to the CATPCA in terms of sharing the percent of the variance. If the variable shares less percent of the variance, it contributes less and vice versa, showing how beneficial the variables are to CATPCA. Suppose it is required to keep a certain set of variables in the analysis. In that case, most of the components with their weights are to be extracted so that the communality for each variable exceeds the minimum threshold of communality significance, i.e., 50%. It can be calculated as the sum of squared component weights for a given variable [15].

Scree Plot Criterion: The scree plot criterion has been used to extract the maximum number of components to work with. A Scree plot is a graphical representation of the eigenvalues against the component number and is very helpful in finding several components for further analysis. It always starts with a high value along the y-axis as it represents the eigenvalue for the first principal component explaining much of the shared variance. Later on, the line starts to dip along the x-axis as the eigenvalues for the rest of the principal components share a lesser and lesser percentage of the variance. The significant knee of the line in two dimensions shows the number of principal components to be selected [22].

Results and Discussion

CATPCA was applied to the obese sample, which started with 0 iterations. The accounted variance of 87.045800 for all the variables at 0 iterations was achieved. Table 2 further shows the iteration history of the CATPCA process. As depicted, the iterations stopped with an accounted variance of 87.108862 for all the variables at 100 iterations.

Table 2: Iteration History ~ Train Data Set for Obese Sample.

Iteration Number	Variance Accounted For		Loss		
	Total	Increase	Total	Centroid Coordinates	Restriction of Centroid to Vector Coordinates
0[a]	87.0458	0.000004	7308.954	7242.085	66.86898
1	87.04636	0.00056	7308.954	7242.085	66.86842
2	87.04711	0.00075	7308.953	7242.084	66.86881
Rows truncated					
98	87.10826	0.000308	7308.892	7241.926	66.96543
99	87.10856	0.000304	7308.891	7241.926	66.96586
100b	87.10886	0.0003	7308.891	7241.925	66.96628

Note: a. Iteration 0 displays the statistics of the solution with all variables, except variables with optimal scaling level Multiple Nominal, treated as numerical.

b. The iteration process stopped because the maximum number of iterations was reached.

Table 3: Model Summary ~ Train Data Set for Obese Sample.

Dim	Cronbach' s Alpha	Variance Accounted For					
		Total (Eigenvalue)	% of variance	Cum %	Eigenval Criterion	Proportion Variance Explained	Scree Plot
1	0.891	8.372	9.74%	9.74%	9.74%	9.74%	9.74%
2	0.742	3.756	4.37%	14.1%	14.10%	14.10%	14.10%
3	0.702	3.261	3.79%	17.89%	17.89%	17.89%	
4	0.678	3.034	3.53%	21.42%	21.42%	21.42%	
5	0.659	2.866	3.33%	24.75%	24.75%	24.75%	
6	0.627	2.63	3.06%	27.81%	27.81%	27.81%	
7	0.608	2.505	2.91%	30.73%	30.73%	30.73%	
8	0.596	2.431	2.83%	33.55%	33.55%	33.55%	
9	0.548	2.181	2.54%	36.09%	36.09%	36.09%	
10	0.517	2.045	2.38%	38.47%	38.47%	38.47%	
11	0.504	1.992	2.32%	40.78%	40.78%	40.78%	
12	0.482	1.908	2.22%	43.00%	43.00%	43.00%	
13	0.476	1.89	2.20%	45.20%	45.20%	45.20%	
14	0.427	1.73	2.01%	47.21%	47.21%	47.21%	
15	0.384	1.612	1.87%	49.08%	49.08%	49.08%	
16	0.37	1.576	1.83%	50.92%	50.92%	50.92%	
17	0.336	1.497	1.74%	52.66%	52.66%	52.66%	
18	0.317	1.457	1.69%	54.35%	54.35%	54.35%	
19	0.31	1.442	1.68%	56.03%	56.03%	56.03%	
20	0.3	1.421	1.65%	57.68%	57.68%	57.68%	
21	0.274	1.372	1.59%	59.28%	59.28%	59.28%	
22	0.251	1.33	1.55%	60.82%	60.82%	60.82%	
23	0.24	1.311	1.52%	62.35%	62.35%	62.35%	
24	0.213	1.267	1.47%	63.82%	63.82%	63.82%	

25	0.197	1.242	1.44%	65.26%	65.26%	65.26%	
26	0.153	1.178	1.37%	66.63%	66.63%	66.63%	
27	0.118	1.132	1.32%	67.95%	67.95%	67.95%	
28	0.068	1.072	1.25%	69.20%	69.20%	69.20%	
29	0.066	1.069	1.24%	70.44%	70.44%	70.44%	
30	0.043	1.045	1.21%	71.66%	71.66%	71.66%	
31	0.015	1.015	1.18%	72.84%	72.84%	72.84%	
32	-0.04	0.962	1.12%	73.95%	73.95%	73.95%	
33	-0.044	0.958	1.11%	75.07%	75.07%	75.07%	
34	-0.07	0.935	1.09%	76.16%	76.16%	76.16%	
35	-0.11	0.902	1.05%	77.21%	77.21%	77.21%	
36	-0.138	0.88	1.02%	78.23%	78.23%	78.23%	
37	-0.151	0.87	1.01%	79.24%	79.24%	79.24%	
38	-0.18	0.849	0.99%	80.23%	80.23%	80.23%	
39	-0.236	0.811	0.94%	81.17%		81.17%	
40	-0.239	0.809	0.94%	82.11%		82.11%	
41	-0.285	0.78	0.91%	83.02%		83.02%	
42	-0.355	0.74	0.86%	83.88%		83.88%	
43	-0.376	0.729	0.85%	84.73%		84.73%	
44	-0.408	0.713	0.83%	85.56%		85.56%	
45	-0.437	0.698	0.81%	86.37%			
46	-0.462	0.687	0.80%	87.17%			
. Rows truncated . .							
85	-226.81	0.004	0.00%	101.29%			
86	0	0	0.00%	101.29%			
Total	1.000a	87.109					

Note: a. Total Cronbach's Alpha is based on total Eigenvalue.

Principal Component Selection Criteria

As depicted in Table 2, the CATPCA algorithm finished iteration, and the eigenvalues were generated for all the 86 principal components with accounted variance shared by each of them. As noted in Table 3, the dimensions that share the maximum percent of the variance were selected. The first dimension shown in the table had an eigenvalue of 8.372, and it shares 9.74% of the total variance, which happens to be the highest percentage of shared variance among all the PCs, the eigenvalue was not very high, and that's because of the greater number of dimensions. Following the first dimension, the second- dimension shares 3.756% of the total variance similarly fifth until tenth and eleventh until thirty- first shares almost the same percentage of the total variance, i.e., ≥ 2 and ≥ 1 respectively. Since the percentage was getting lower than 1%, choosing dimensions was obvious, i.e., 31 dimensions. The next seven eigenvalues for the principal components were not very far from the threshold value of 1, so these components were also included. These dimensions shared 80.23% ≈ and 80% percent of the total variance, which was not as good as required, but this was the maximum number of dimensions best suited for this data set. The relevant criterion to extract the principal components was checked and finalized in the next section.

Component Extraction Criterion

As discussed in Sections 2.4.1 to 2.4.2.4, the criterion was to be applied to the results to finalize the PCs so that profiling can be processed to know the characteristics of these PCs, respectively. The results presented by the algorithm in Table 3 give us the model summary for the percent of variance shared by all the PCs. Based on these eigenvalues, it was further evaluated to suggest and extract the number of dimensions and PCs. Further presented in the table are the 86 dimensions for 100% of the shared variance in the data set. These dimensions were evaluated with criteria, and then the profiling of these PCs was finalized. The Eigenvalue criterion selected

thirty-eight dimensions sharing an approximate 80.23% ≈ 80% of the total variance, which supports the theory of its tendency to extract more dimensions of variables in the dataset are > 50 variables. The proportion of variance explained criterion selected 44 dimensions sharing an approximate 85.56% ≈ 86% of the total variance. The knee of the scree plot depicted in Figure 3 suggested two principal components sharing an approximate 14.10% ≈ and 14% of the total variance to work with. Since all the criteria were analyzed and applied to extract the exact number of components, finally, it was agreed to apply the eigenvalue criterion to extract the number of principal components. It was due to consideration of the said criterion for the variables that had eigenvalue ≥ 1 as it defines the one's variable worth; it also extracted a lesser number of principal components (38) comparatively with a reasonable percent of shared variance among the PCs, i.e., ≈ 80%. A total of 38 principal components reduced from 86 principal components were considered, as shown in Table 3. Before profiling the principal components, the component weights also had to be evaluated.

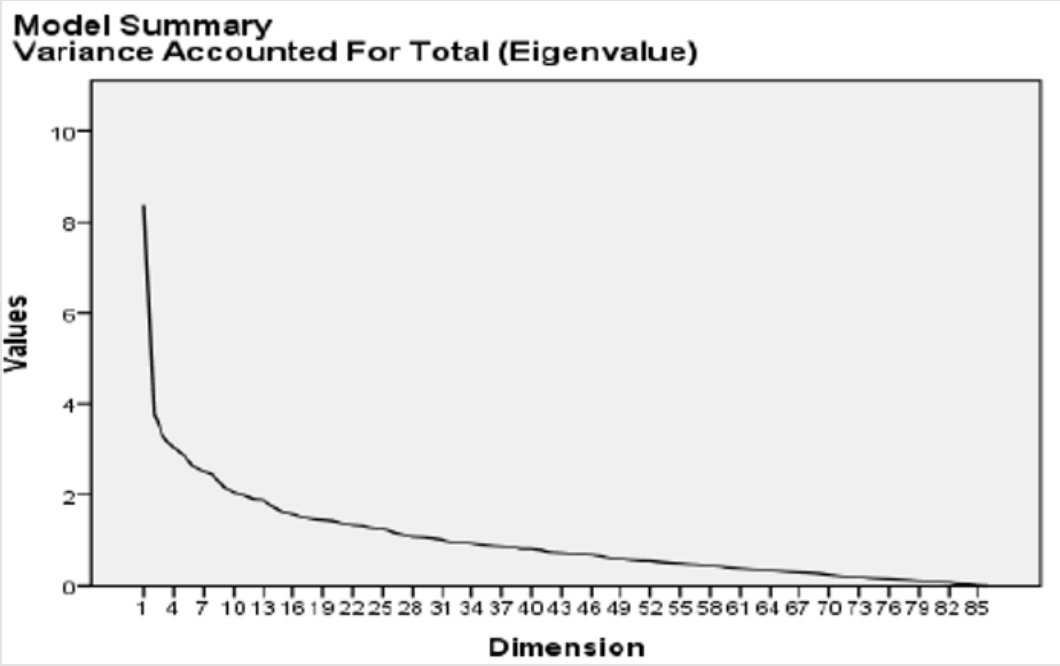

Figure 3: Scree Plot Criterion for Component Selection ~ Obese Sample.

Once the dimensions were chosen for the shared percent of variance by the principal components, it was time to evaluate and extract the components based on their factor weights. For evaluation of the component weights, the weight threshold value equal to +/- 0.50 was considered to retain the component, which would define its contribution to CATPCA as a whole. The components that had component weight less than +/-0.50 were to be excluded as the decided threshold value in this research was +/-0.50 or values close to it. Finally, 28 principal components were further excluded based on criterion and component weights, with ten principal components retained for profiling. The principal components extracted were PC6, 10-12, 14-26, and 28-38. An overview of the extracted PCs concerning the factors within respective PCs has been presented in Figure 4. The figure shows all factors within PC along the y-axis, while along the x-axis are extracted PCs. As discussed in Section 3.1, it is noticeable that PC1 has most of the variables count (PC1 = 16 variables), which correlates with the percent of variance shared among the PCs as normally, the first PC has the maximum percent of the shared the factors allotted to respective PCs. Later on, all the principal components are discussed to note the salient features of all the PCs, respectively.

Principal Component 1: PC1 presented in Table 4 is composed largely of the "block group size" variables, namely, history of raised blood pressure: tablets taken (53a), diet (53b), lose weight (53c), stop smoking (53d), start exercise (53e), history of diabetes: blood sugar measured in past 12 months (56), tablets taken (60b), diet (60c), weight loss (60d), smoking (60e), start exercise (60f), history of raised blood cholesterol: tablets taken (65a), diet (65b), weight loss (65c), stop smoking (65d), and start exercise (65e) all have large values referred to as high levels. The values presented in Table 4 show that these variables were right skewed. It means most individuals were not receiving any advice from the doctor or treatment in terms of tablets, prescribed diet plan, weight loss, stop smoking habit, start or stop the exercise as far as the history of raised blood pressure was concerned. The same trend was observed for a history of diabetes and a history of high blood cholesterol. PC1 shares the maximum percentage of variance by factors. The salient characteristics showed that this component belonged to healthy individuals with no history of raised blood pressure, diabetes, and high blood cholesterol.

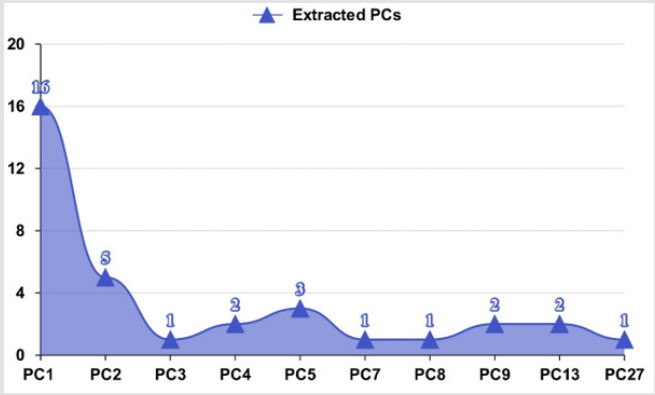

Figure 4: CATPCA Final Extracted Components.

Table 4: Components Extraction ~ CATPCA.

Sr.#	Variables	PC. No.	PC Wts	Sr. #	Variables	PC. No.	PC Wts
1	ageyears	PC3	0.511091	18	Diet60c	PC1	0.669637
2	DistCd	PC4	-0.48467	19	Wtls60d	PC1	0.760748
3	Gndr6	PC2	-0.77543	20	Smk60e	PC1	0.637105
4	Relgn10	PC7	-0.51971	21	Exer60f	PC1	0.760748
5	Elec13	PC13	0.515694	22	HadCh64	PC5	0.600427
6	Pwtr14	PC13	0.55219	23	Tab65a	PC1	0.634584
7	Cusmk24	PC2	-0.52314	24	Diet65b	PC1	0.64151
8	Smkls33	PC4	0.52311	25	Lswt65c	PC1	0.790191
9	HadBP52	PC5	0.531008	26	Smk65d	PC1	0.623979
10	Tab53a	PC1	0.587767	27	Anem68f	PC27	-0.4748
11	Diet53b	PC1	0.516721	28	Imge69	PC9	0.545242
12	Lswt53c	PC1	0.657802	29	WtNow70	PC9	0.533603
13	Smk53d	PC1	0.569552	30	Wt73	PC2	0.696849
14	Exer53e	PC1	0.657887	31	Ht74	PC2	0.783205
15	BdSgr56	PC1	0.537193	32	Wst75	PC2	0.599905
16	HadDM57	PC5	0.610818	33	NasiK90	PC8	-0.4644
17	Tab60b	PC1	0.577	34	Exer65e	PC1	0.782187

Principal Component 2: Table 4 depicts PC2, which is about demographic status, Socio-Economic status, smoking status, recreational activity, and body image. It showed that most of the individuals were female (6). Mostly never smoked any tobacco products such as cigarettes, cigars, or pipes in recent times (24). The body image showed that the individuals in this principal component had heavyweights (73) and heights (74) along with their waists (75).

Principal Component 3: PC3 presents the demographic characteristics of the sample as presented in Table 4. The age in years and values were noted high and increasing concerning obesity which means this variable had contributed well to CATPCA. Most of the individuals were elderly aged (1).

Principal Component 4: PC4 presents the sample's demographic, socio- economic, smoking, and health characteristics in Table 4. The level states that most individuals lived in the main districts (DistCd) of Brunei Darussalam. They were Brunei Citizens (8) and did not smoke daily (33) tobacco products such as cigarettes, cigars, or pipes.

Principal Component 5: PC5 in Table 4 presents physical activity status, history of raised blood pressure, and blood cholesterol. It showed high values for all the factors. The individuals in this PC did not know about being told by a doctor or health worker about having high bp or hypertension (52); similarly, they had never been told by a doctor or health worker of high blood sugar levels or diabetes (57) during the past 12 months. They also had never been told by a doctor or health worker about high blood cholesterol (64) during the past 12 months.

Principal Component 7, 8, 9: PC7, 8, and 9 in Table 4 present the individuals' demographic status, body image, and short food frequency status. PC7 depicts that most individuals were Muslim belonging to the religion (10) Islam. PC8 & PC9 in Table 4 presented

the body image and short food frequency status of the obese sample. Most of the individuals considered themselves Overweight (69) and were not satisfied with their body weights (70), while most of the people were used to eating nasi katok (90) and Chicken Tail / Wings / Skin (91) twice a week.

Table 5: Components Loadings ~ Train Data Set for Obese Sample.

Variables	Dimension									
	1	2	3	4	5	7	8	9	13	27
Age years	-0.382	-0.158	0.511	-0.079	-0.168	0.087	0.127	-0.032	-0.053	0.045
DistCd	0.047	0.057	0.221	-0.485	0.394	-0.335	-0.008	0.038	0.052	0.133
Gndr6	0.222	-0.775	0.016	-0.124	-0.04	-0.108	0.143	-0.039	-0.013	-0.003
Relgn10	0.056	0.201	0.154	-0.153	0.106	-0.52	-0.128	0.189	-0.221	0.085
Elec13	0.054	-0.006	0.072	-0.118	-0.044	-0.316	0.139	0.196	0.516	0.025
Pwtr14	0.03	0.03	0.026	-0.005	-0.022	-0.22	0.187	0.132	0.552	-0.019
Cusmk24	0.075	-0.523	0.116	0.11	0.205	-0.121	0.003	-0.04	-0.06	0.02
Smkls33	-0.024	-0.082	-0.087	0.523	0.362	0.054	0.485	0.25	-0.042	0.023
HadBP52	0.038	-0.013	-0.033	-0.375	0.531	0.341	-0.004	0.072	0.015	-0.069
Tab53a	0.588	0.14	-0.395	-0.101	-0.131	0.013	-0.026	0.061	0.087	0.017
Diet53b	0.517	0.198	-0.377	-0.147	-0.176	-0.005	-0.011	0.15	0.049	-0.02
Lswt53c	0.658	0.045	-0.342	-0.094	-0.053	-0.149	0.032	-0.073	0.125	0.03
Smk53d	0.57	-0.198	0.011	0.193	0.165	-0.056	0	-0.082	0.02	-0.084
Exer53e	0.658	0.061	-0.383	-0.1	-0.066	-0.122	-0.001	-0.039	0.145	0.017
BdSgr56	0.537	0.185	-0.14	-0.03	0.074	0.007	0.019	0.116	0.048	0.003
HadDM57	0.092	0.041	0.012	-0.443	0.611	0.327	0	0.102	0.042	-0.001
Tab60b	0.577	0.232	0.374	0.074	-0.11	0.307	0.063	0.039	0.119	0.018
Diet60c	0.67	0.197	0.37	0.061	-0.108	0.249	0.048	0.071	0.059	0.005
Wtls60d	0.761	0.121	0.383	0.126	-0.015	0.125	-0.007	-0.024	0.113	0.026
Smk60e	0.637	-0.121	0.361	0.15	0.172	-0.046	-0.141	-0.052	-0.008	-0.043
Exer60f	0.761	0.121	0.383	0.126	-0.015	0.125	-0.007	-0.024	0.113	0.026
HadCh64	0.087	0.058	0.01	-0.416	0.6	0.341	0.012	0.08	0.046	-0.032
Tab65a	0.635	0.158	0.08	-0.007	-0.137	0.077	0.131	-0.023	-0.15	-0.05
Diet65b	0.642	0.165	0.016	-0.066	-0.15	-0.034	0.194	0.042	-0.193	-0.013
Lswt65c	0.79	0.027	0.084	-0.001	-0.042	-0.124	0.125	-0.103	-0.164	0.001
Smk65d	0.624	-0.178	0.251	0.126	0.17	-0.133	-0.093	-0.076	-0.113	-0.075
Anem68f	-0.031	0.137	-0.123	0.053	-0.008	0.07	-0.151	0.198	0.058	0.475
Imge69	0.058	-0.077	-0.122	0.105	-0.047	-0.092	0.064	0.545	-0.143	0.021
WtNow70	0.075	-0.086	-0.164	0.112	-0.046	-0.096	0.067	0.534	-0.13	0.029
Wt73	-0.34	0.697	0.082	0.14	0.035	-0.049	-0.019	0.033	-0.134	-0.041
Ht74	-0.234	0.783	0.125	0.147	0.111	0.014	-0.156	0.032	-0.033	-0.029
Wst75	-0.439	0.6	0.063	0.066	-0.023	0.008	0.026	0.077	-0.167	-0.056
NasiK90	0.006	-0.147	0.21	0.101	0.101	0.039	-0.464	0.212	-0.152	-0.156
Exer65e	0.782	0.043	0.069	-0.025	-0.041	-0.142	0.142	-0.108	-0.181	-0.017

Principal Component 13: PC13 in Table 4 depicts Socio-Economic status with high values. It depicts that most individuals had electricity and water piped supply (13 and 14) to their houses.

Principal Component 27: PC27 in Table 4 represents the health status which showed that most of the individuals were suffering from anemia (68f) as far as health was concerned.

Minimum Communalities Criterion

As discussed in Section 2.4.2.3 and Table 3, the variable that shares less communality means shares less of its common variability among the variables, and contribution to the CATPCA is also considered lesser. At first, the finalized PCs were compiled concerning the factor variable weights (≥ +/-0.50), as highlighted in Table 5. The communality values showed the contributing factor variables. 35-factor variables out of the total 86 variables met the criteria, and the rest were omitted. In the second step, all the respective PCs' weights

Table 6: Squared Components Loadings Communality ~ Train Data Set for Obese Sample.

Variables	Dimension									Comm. Criteria	
	1	2	3	4	5	7	8	9	13	27	Sqrd.Sum
Age years	0.146	0.025	0.261	0.006	0.028	0.008	0.016	0.001	0.003	0.002	0.497
DistCd	0.002	0.003	0.049	0.235	0.155	0.112	0	0.001	0.003	0.018	0.579
Gndr6	0.049	0.601	0	0.015	0.002	0.012	0.021	0.002	0	0	0.702
Relgn10	0.003	0.04	0.024	0.023	0.011	0.27	0.016	0.036	0.049	0.007	0.48
Elec13	0.003	0	0.005	0.014	0.002	0.1	0.019	0.038	0.266	0.001	0.448
Pwtr14	0.001	0.001	0.001	0	0	0.048	0.035	0.017	0.305	0	0.409
Cusmk24	0.006	0.274	0.014	0.012	0.042	0.015	0	0.002	0.004	0	0.367
Smkls33	0.001	0.007	0.008	0.274	0.131	0.003	0.235	0.062	0.002	0.001	0.722
HadBP52	0.001	0	0.001	0.141	0.282	0.116	0	0.005	0	0.005	0.552
Tab53a	0.345	0.019	0.156	0.01	0.017	0	0.001	0.004	0.007	0	0.561
Diet53b	0.267	0.039	0.142	0.021	0.031	0	0	0.022	0.002	0	0.526
Lswt53c	0.433	0.002	0.117	0.009	0.003	0.022	0.001	0.005	0.016	0.001	0.608
Smk53d	0.324	0.039	0	0.037	0.027	0.003	0	0.007	0	0.007	0.446
Exer53e	0.433	0.004	0.147	0.01	0.004	0.015	0	0.002	0.021	0	0.635
BdSgr56	0.289	0.034	0.02	0.001	0.005	0	0	0.013	0.002	0	0.365
HadDM57	0.008	0.002	0	0.196	0.373	0.107	0	0.01	0.002	0	0.699
Tab60b	0.333	0.054	0.14	0.005	0.012	0.094	0.004	0.002	0.014	0	0.659
Diet60c	0.448	0.039	0.137	0.004	0.012	0.062	0.002	0.005	0.003	0	0.712
Wtls60d	0.579	0.015	0.147	0.016	0	0.016	0	0.001	0.013	0.001	0.786
Smk60e	0.406	0.015	0.13	0.022	0.03	0.002	0.02	0.003	0	0.002	0.63
Exer60f	0.579	0.015	0.147	0.016	0	0.016	0	0.001	0.013	0.001	0.786
HadCh64	0.008	0.003	0	0.173	0.361	0.116	0	0.006	0.002	0.001	0.671
Tab65a	0.403	0.025	0.006	0	0.019	0.006	0.017	0.001	0.023	0.003	0.502
Diet65b	0.412	0.027	0	0.004	0.022	0.001	0.037	0.002	0.037	0	0.544
Lswt65c	0.624	0.001	0.007	0	0.002	0.015	0.016	0.011	0.027	0	0.702
Smk65d	0.389	0.032	0.063	0.016	0.029	0.018	0.009	0.006	0.013	0.006	0.579
Exer65e	0.612	0.002	0.005	0.001	0.002	0.02	0.02	0.012	0.033	0	0.706
Anem68f	0.001	0.019	0.015	0.003	0	0.005	0.023	0.039	0.003	0.225	0.334
Imge69	0.003	0.006	0.015	0.011	0.002	0.009	0.004	0.297	0.02	0	0.368
WtNow70	0.006	0.007	0.027	0.013	0.002	0.009	0.004	0.285	0.017	0.001	0.371
Wt73	0.115	0.486	0.007	0.02	0.001	0.002	0	0.001	0.018	0.002	0.652
Ht74	0.055	0.613	0.016	0.022	0.012	0	0.024	0.001	0.001	0.001	0.745
Wst75	0.193	0.36	0.004	0.004	0.001	0	0.001	0.006	0.028	0.003	0.599
NasiK90	0	0.022	0.044	0.01	0.01	0.002	0.216	0.045	0.023	0.024	0.396

(Table 5) according to the communality criterion were calculated with their squared weights. Table 6 depicts the squared weights for all the factor variables in respective PCs. It shows the squared component loadings for the 25-factor variables that met the criteria by having a communality significance value ≥ 50% showing their contribution to CATPCA. It means these are the final set of factor variables that have contributed well to the algorithm as a whole. CATPCA classified the NHANSS data into two subgroups; one subgroup was presented with left-skewed distribution while the other was presented with right-skewed distribution, which means that the most prevalent conditions concerning obesity by variables were either detected or undetected. The variables that presented the communality significance more than the threshold (50%) value were the variables that helped gain knowledge about the salient characteristics of the NHANSS obese sample. As discussed above, the generic details of these variables are below mentioned for reference.

1) Demographic and Socio-Economic Characteristics
2) Smoking Characteristics
3) History of Raised Blood Pressure
4) History of Diabetes Mellitus
5) History of High Blood Cholesterol and,
6) Anthropometric Characteristics

Table 7: Model Summary for Test Data Set ~ Obese Sample.

Dim	Cronbach' s Alpha	Variance Accounted For					
		Total (Eigenval)	% of variance	Cum %	Eigenvalue Criterion	Proportion Variance Explained	Scree Plot
1	0.91	9.954	11.44%	11.44%	11.44%	11.44%	11.44%
2	0.783	4.421	5.08%	16.52%	16.52%	16.52%	16.52%
3	0.763	4.057	4.66%	21.19%	21.19%	21.19%	
4	0.702	3.262	3.75%	24.94%	24.94%	24.94%	
5	0.668	2.94	3.38%	28.32%	28.32%	28.32%	
6	0.657	2.851	3.28%	31.59%	31.59%	31.59%	
7	0.65	2.8	3.22%	34.81%	34.81%	34.81%	
8	0.641	2.727	3.13%	37.95%	37.95%	37.95%	
9	0.616	2.559	2.94%	40.89%	40.89%	40.89%	
10	0.577	2.327	2.68%	43.56%	43.56%	43.56%	
11	0.541	2.151	2.47%	46.03%	46.03%	46.03%	
12	0.521	2.061	2.37%	48.40%	48.40%	48.40%	
13	0.518	2.048	2.35%	50.76%	50.76%	50.76%	
14	0.493	1.951	2.24%	53.00%	53.00%	53.00%	
15	0.451	1.801	2.07%	55.07%	55.07%	55.07%	
16	0.439	1.765	2.03%	57.10%	57.10%	57.10%	
17	0.408	1.675	1.93%	59.02%	59.02%	59.02%	
18	0.387	1.618	1.86%	60.88%	60.88%	60.88%	
19	0.368	1.573	1.81%	62.69%	62.69%	62.69%	
20	0.334	1.493	1.72%	64.41%	64.41%	64.41%	
21	0.305	1.433	1.65%	66.05%	66.05%	66.05%	
22	0.268	1.36	1.56%	67.62%	67.62%	67.62%	
23	0.242	1.313	1.51%	69.13%	69.13%	69.13%	
24	0.225	1.286	1.48%	70.61%	70.61%	70.61%	
25	0.205	1.252	1.44%	72.04%	72.04%	72.04%	
26	0.174	1.209	1.39%	73.43%	73.43%	73.43%	
27	0.152	1.178	1.35%	74.79%	74.79%	74.79%	
28	0.09	1.099	1.26%	76.05%	76.05%	76.05%	
29	0.062	1.066	1.23%	77.28%	77.28%	77.28%	
30	0.046	1.049	1.21%	78.48%	78.48%	78.48%	
31	0	1	1.15%	79.63%	79.63%	79.63%	
32	-0.04	0.962	1.11%	80.74%	80.74%	80.74%	
33	-0.095	0.914	1.05%	81.79%	81.79%	81.79%	
34	-0.151	0.87	1.00%	82.79%	82.79%	82.79%	
35	-0.18	0.848	0.97%	83.76%		83.76%	
36	-0.203	0.833	0.96%	84.72%		84.72%	
37	-0.242	0.808	0.93%	85.65%			
38	-0.323	0.758	0.87%	86.52%			
. Rows truncated . .							
87	-86.921	0.012	0.01%	89.71%			
Total	1.000a	87.999					

Note: a. Total Cronbach's Alpha is based on total Eigenvalue.

Validation of Principal Components ~ Obese Sample

As discussed in Section 2.3, the test data set taken from 449 instances (obese sample) was divided with a ratio (70:30) of 314:135, respectively. CATPCA generated the results on the test data set, and then these results were compared for validation of principal components with those already generated aforementioned. It was noticeable that the results generated by the train data set did not show much difference concerning the selection and extraction of components for further evaluation. The process started with 0 iterations and ended at 100 iterations. As shown in Table 7, the shared variance was noted 87.999 as a whole by the CATPCA. The model summary was generated against the eigenvalues representing the percent of variance shared among the principal components. To evaluate these PCs and to know whether this test data set had also generated the same number of PCs, a comparison had to be made to indicate whether these results for the data set as a whole are generalized or not, so the results can be reported as Valid or Invalid. The eigenvalues starting from PC1, both the data sets, train, and test data sets, almost shared the same percentage of variance reported as 8.372 and 9.954, respectively. Similarly, for PC2, the eigenvalues were reported as 3.756 and 4.421, respectively. For PC3, the eigenvalues were reported as 3.261 and 4.057, respectively, and so on. Here, it is wise to compare the criterion results from train and test data sets to see if the reported results were the same as those of eigenvalues or if they differ significantly. If there were a minimal difference in the number of selected PCs or shared variance, it would validate the results, but if vice versa, then the validation would be reported as invalid as far as the reporting of the results was concerned. Since the eigenvalue criteria were finalized for the train data set, the results concerning the eigenvalue criterion generated by the test data set were checked and compared for validation.

Eigen Value Criterion ~ Test Data Set: The results in Table 7 showed the same trend of extracting more PCs in terms of dimensions as the data set had more than 50 variables. Hence, the criterion suggests extracting exactly 31 dimensions with eigenvalue ≥ 1. The next three proceeding dimensions with eigenvalues close to 1, i.e., ≤ 0.85, were added. A total of 34 dimensions were suggested by this criterion, sharing approximately 82.79% ≈ 83% of the total variance, which again supported the theory of its tendency to extract more dimensions (if variables in the data set are > 50 variables). Comparing it to the eigenvalue criterion results generated by the test data set seems to validate the results generated by the train data set, as discussed in Section 3.3. The eigenvalue criterion on the test data set suggested 34 dimensions with an estimated shared percent variance of 83%, which validates the eigenvalue criterion results generated by the train data set (the suggested result was 38 dimensions with an estimated shared percent variance of 80%). The results did not show any huge difference in the dimensions' shared percent variance, and almost the same number of dimensions were selected. It shows that these details validate the results generated by the train data set and now can be reported as Valid.

Conclusion

Obesity is one of the non-communicable diseases that is a condition of being overweight or a major nutritional disorder. The prevalence of obesity in Brunei Darussalam has increased more than double since 1997, to 27.2% in 2011, and around 61% of Bruneians are overweight and obese, which is highest in the ASEAN region. Comparatively, in the US, the prevalence of obesity in 2011-2014 was 22.8% (including obese and extremely obese individuals) among the youth aged 2-19 years which shows that obesity has become a worldwide epidemic. Its growth has been projected at 40% in the upcoming decade. In this study, the classification technique was used to identify the obesity subgroups within the NHANSS data provided by the ministry of health, Brunei Darussalam. The novelty of the research was to extract useful knowledge from NHANSS data of mixed variable types as not many studies have been carried out in the past in this domain with mixed data types. CATPCA algorithm was used, which grouped the obese sample into two classes concerning the anchoring conditions related to obesity. The two subgroups presented the most prevalent conditions belonging to demographic, Socio-Economic, smoking, anthropometric, and short food frequency characteristics of the obese sample. The short food frequency revealed that the obese group was not taking care of their diet and was used to eating nasi katok (local rice cooked with fried chicken) and chicken Tail / Wings / Skin twice a week. Noticeably the history of blood pressure, diabetes mellitus, and high blood cholesterol were undetected for obese patients, but most of them were reported as having anemia as far as their health was concerned. All of these results were validated, and profiling was noted accordingly. This research is of clinical importance, and the salient features should be reported and further investigated from a medical perspective. The proposed approach reveals the sub-groups that may help investigate the importance of the lifestyle factors (i.e., age, smoking habits, blood pressure, diabetes mellitus, high blood cholesterol, etc.) from a clinical point of view. Overall, the combination of clinical knowledge with data-hidden information and the evaluation of subclasses revealed by the data structure could lead to very interesting developments.

Acknowledgments

The authors would like to express sincere appreciation for the technical assistance and support from the Department of Economic Planning and Development Brunei Darussalam, research assistant, and managers from the Ministry of Health Brunei Darussalam, and participation from the survey respondents.

Conflict of Interest

The author(s) declared no potential conflicts of interest concerning this article's research, authorship, and/or publication.

References

1. Brunei (2005) Brunei Darussalam Government Gazette Part iii Smoking in Specified Places and Specified Vehicles. Bandar Seri Begawan, Brunei Darussalam.
2. L Uccioli, G Monticone, F Russo, F Mormile, L Durola, et al. (1994)

Autonomic neuropathy and transcutaneous oxymetry in diabetic lower extremities. Diabetologia 37(10):1051-1055.

3. (2013) Ministry of Health Brunei Darussalam, Brunei Darussalam National Multisectoral Action Plan for the Prevention and Control of Noncommunicable Diseases 2013-2018, Bandar Seri Begawan, Brunei Darussalam.
4. Hanafi (2017) Message by Yang Berhormat Dato Seri Setia Dr Haji Zulkarnain Bin Haji Hanafi Minister Of Health On The Occasion Of World Cancer Day 2017, Moh.Gov.Bn.
5. IA WM Nazlee WZ, Rosnani Z (2019) Brunei International. Brunei Int Med J 15: 53-57.
6. (2018) I ASEAN Secretariat, Jakarta, The ASEAN Secretariat Jakarta. Jakarta, Indonesia: ASEAN Secretariat, December 2018.
7. Sok King Ong, Daphne Teck Ching Lai, Justin Yun Yaw Wong, Khairil Azhar Si-Ramlee, Lubna Abdul Razak, et al. (2017) Cross-sectional STEPwise Approach to Surveillance (STEPS) Population Survey of Noncommunicable Diseases (NCDs) and Risk Factors in Brunei Darussalam 2016. Asia-Pacific J Public Heal 29(8): 635-648.
8. U Khalil, OA Malik, D Lai, OS King (2018) Identifying sub-groups of the obese from national health and nutritional status survey data using machine learning techniques," in IET Conference Publications. CP750, 113 (4 pp.)-113 (4 pp.).
9. CL Ogden, MD Carroll, BK Kit, M Flegal, Ogden CL, et al. (2016) Prevalence of Childhood and Adult Obesity in the United States, 2011-2012. Jama 311(8): 806-814.
10. (2014) ASEAN Secretariat, Association of Southeast Asian Nations, Annual Report, 2013-2014. Jakarta, Indonesia: JAKARATA, ASEAN Secretariat.
11. B MoH (2014) The Report, The 2nd National Health and Nutritional Status Survey (NHNANSS) 2014. Ministry of Health, Commonwealth Drive, Brunei Darussalam, Bandar Seri Begawan.
12. A Othman (2020) Brunei records highest child obesity rate in region | Borneo Bulletin Online, Borneo Bulletin.
13. U Khalil, OA Malik, D Teck, C Lai, OS King (2021) Cluster Aanalysis for Identifying Obesity Subgroups in Health and Nutritional Status Survey Data. Asia-Pacific J Inf Techno Multimed 10(2): 146-169.
14. N Antonioli, F Castanò, S Coletta, S Grossi, Domenico Lembo, et al. (2014) Ontology-based data management for the Italian public debt. Frontiers in Artificial Intelligence and Applications 267: 372-385.
15. A Ghatak (2017) Machine Learning with R (2nd Edn.).,. Livery Place 35 Livery Street Birmingham B3 2PB, UK.
16. MA Green, M Strong, F Razak, SV Subramanian, C Relton, et al. (2016) Who are the obese? A cluster analysis exploring subgroups of the obese. J Public Heal (United Kingdom) 38(2): 258-264.
17. I Kavakiotis, O Tsave, A Salifoglou, N Maglaveras, I Vlahavas, et al. (2017) Machine Learning and Data Mining Methods in Diabetes Research. Comput Struct Biotechnol J 15: 104-116.
18. Y Mori, M Kuroda, N Makino (2016) Nonlinear Principal Component Analysis and Its Applications springer briefs in statistics. Springer briefs Stat (1): 2-85.
19. CA Befort, N Nazir, MG Perri (2012) Behavior Risk Factor Surveillance System (BRFSS) 5 and the 1997-1998 National Health J Rural Health J Rural Heal 28(4): 392-397.
20. J won Lee, C Giraud-Carrier (2013) Results onmining NHANES data: A case study inevidence-based medicine. Comput Biol Med 43(5): 493-503.
21. Linting M, Meulman JJ, Groenen PJF, Van der Kooij (2004) Nonlinear Principal Components Analysis. Am Psychol Assoc, p. 12-48.
22. M Linting, JJ Meulman, PJF Groenen, AJ Van der Kooij (2007) Nonlinear Principal Components Analysis: Introduction and Application. Psychol Methods 12(3): 336-358.

ISSN: 2574-1241

DOI: 10.26717/BJSTR.2023.48.007641

Usman Khalil. Biomed J Sci & Tech Res