5-2023

A Machine Learning Approach to Obese-Inflammatory Phenotyping

Tania Mayleth Vargas
The University of Texas Rio Grande Valley

A MACHINE LEARNING APPROACH TO OBESE-INFLAMMATORY PHENOTYPING

A Thesis

by

TANIA MAYLETH VARGAS

Submitted in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Major Subject: Applied Statistics and Data Science

The University of Texas Rio Grande Valley

May 2023

A MACHINE LEARNING APPROACH TO OBESE-INFLAMMATORY PHENOTYPING

A Thesis
by
TANIA MAYLETH VARGAS

COMMITTEE MEMBERS

Dr. Kristina Vatcheva
Chair of Committee

Dr. Mrinal Roychowdhury
Committee Member

Dr. Santanu Chakraborty
Committee Member

Dr. Xiaohui Wang
Committee Member

May 2023

ABSTRACT

Vargas, Tania M., A Machine Learning Approach to Obese-Inflammatory Phenotyping. Master of Science (MS), May, 2023, 51 pp., 12 tables, 12 figures, references, 67 titles.

Obesity is the accumulation of an abnormal, or excessive, amount of fat in the body, which can have negative effects on overall health. This excess accumulation of macronutrients in adipose tissue can cause the release of inflammatory mediators, leading to a pro-inflammatory state. Inflammation is a known risk factor for various health conditions, including cardiovascular diseases, metabolic syndrome, and diabetes. This study sought to examine the use of data mining methods, particularly clustering algorithms, to identify inflammatory biomarker phenotypes and their association with obesity in a local adolescent population. The algorithms evaluated in this study included: *k*-means, Ward's hierarchical agglomerative method, fuzzy *c*-means, Gaussian mixture model, and principal component analysis (PCA). The algorithms were assessed using different validation indices, graphs, as well as clinical interpretation of the resulting clusters. The results showed that *k*-Means, $k = 3$, produced the most accurate clusters. Based on their characterization, the clusters were defined as: severe risk for metabolic dysfunction, moderate risk for metabolic dysfunction, and normal metabolic function. Adolescents with a higher BMI and waist circumference had higher odds of being classified in the severe metabolic risk cluster. Although PCA is a different type of clustering algorithm, it supported the resultant cluster by grouping their dominant inflammatory biomarkers characteristics into separate principal components.

These findings suggested a strong relationship between CRP and Leptin inflammatory biomarkers and higher BMI and waist circumference in the local adolescent study population.

DEDICATION

To my loving family. Your unwavering love and support have been the driving force behind my achievements. Thank you for always believing in me.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my Committee Chair, Dr. Kristina Vatcheva, for her invaluable guidance, continuous support, and insightful feedback throughout the entire thesis process. Her dedication, expertise, and willingness to share her knowledge has been instrumental in shaping my research work. I am incredibly grateful for her mentorship, patience, and encouragement.

I would like to extend my thanks to Dr. Mrinal Roychowdhury, Dr. Santanu Chakraborty, and Dr. Xiaohui Wang for serving as Committee Members for my thesis. Their time, expertise, and guidance were critical to the completion of my thesis.

A special thank you to Dr. Saraswathy Nair for providing me the data used in my research. I appreciate your support and assistance.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Motivation

In modern society, the collection of data continues to increase. Due to the expansion of technology, we have the capability to not only collect data directly from subject of interest, but also from different data generating sources. However, collected data may be gathered in various forms and thus be different in terms of quality. Raw data can be unorganized and messy. Researchers have begun to use machine learning techniques to gain understanding and learn from the data or simply to organize and store data accurately. There are two main types of machine learning techniques, supervised and unsupervised (Jain et al, 1999). Supervised techniques allow the researchers to discover patterns and variables from the labeled data set. Labeled data are data with known outcome (target) variable(s). Unlabeled data sets are datasets in which there is no particular outcome variable, or classification variable, nor is anything known about the relationship between the observations. When significant characteristics or properties are identified for any variables in an unlabeled data set, then a meaningful label, or class, may be designated to the pertaining variable to convey its significance. The supervised algorithm will learn from the labeled variables in a training data, and use those observations to predict the value, or class, of an unlabeled variable. Unsupervised techniques use machine learning algorithms to learn and discover hidden patterns from unlabeled data. These algorithms group observations that are most similar to each other and maximizes the dissimilarity to the other groups. The creation of different groups allows

the researcher to be able to discern any relationships between the variables by analyzing the patterns (Gentleman et al, 2008). In a data-driven world, these methods have been powerful tools in research.

Methods of data collection may vary, and the obtained data may come in different forms or arrangement. Since the features in the data sets will not always appear clearly categorized into groups or labels, it is crucial to emphasize the importance of unsupervised algorithms when it comes to dealing with unlabeled data (Nwogbaga, 2020). One of the strengths of unsupervised learning machines is their autonomy and independence. However, it is important to note that one of the drawbacks with clustering algorithms in the era of Big Data, is their need for improvement in stability and time. (Zerhari et al, 2015). Although unsupervised learning can answer clustering and association problems, the main goal for the algorithm is to identify hidden patterns among the dataset, group the observations based on a similarity index, and produce a succinct summary without the use of a teacher (Hiran et al, 2021). Algorithms such as Principal Component Analysis (Chi-Hsien & Nagasawa, 2019), Generative Adversarial Networks (McAlpine et al, 2022), and varying clustering methods (Syarif et al, 2012) are a few of the unsupervised learning algorithms that have been implemented for data analysis in different fields. It is important to note that different unsupervised algorithms have different functionality. PCA is a data reduction technique that reduces high-dimensional data into lower dimensions by creating new variables, which are known as the principal components, when combining variables with similar properties or characteristics from the original dataset (Abdi & Williams, 2010). Essentially, PCA allows the researchers to remove redundant variables, or the columns, in a dataset and unveil the core variables that re-express the original high-dimensional dataset in a compressed version (Shlens, 2014). On the contrary, clustering methods cluster observations, or the rows, in a data set based on similarity and

maintain all the original variables. With increasing interest in unsupervised learning machines, clustering algorithms have taken center stage due to their efficiency in grouping data points (i.e., subjects) based on specified properties the algorithm learns (Jain et al, 1999). A few examples of clustering algorithms include *K*-Means, Hierarchical, DBSCAN, Expectation Maximization, and Gaussian Mixture Model. One of the key steps when using these algorithms, is to evaluate the clusters properly to decide which clustering method is most relevant with the input dataset to optimize the results (Palacio-Niño & Berzal, 2019). Figure 1 demonstrates the general process when applying a clustering algorithm.

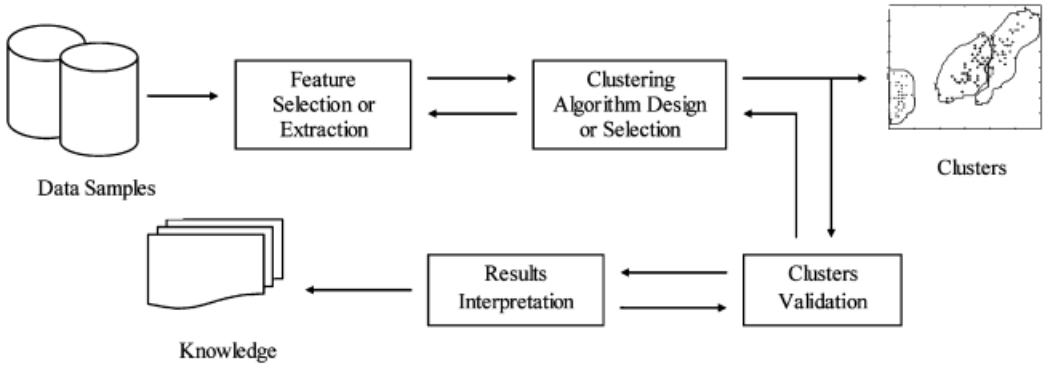

Figure 1 General application of clustering algorithms (Xu & Wunsch, 2005)

Cluster analysis has been widely implemented in the field of academics, data mining, epidemiology, and the marketing field, to name a few. It has been a focal point in some studies to use cluster analysis primarily to recognize patterns and identify possible linkage between variables. For instance, Scherzer et al. (2018) conducted *k*-means clustering on clinical data consisting of 332 HIV positive men and women to determine whether cardiac phenotypes can be identified among the 8 biomarkers being studied. As a result, 3 clusters were produced with different phenotypes. Cluster 1 was the normal group that had the lowest mean levels for all measured biomarkers. Cluster 2 was identified as the cardiac phenotype because it had the highest levels of ST2, NT-proBNP, and GDF-1. Cluster 3 was identified as the inflammatory phenotype because it

had the highest levels of CRP, IL-6, and D-dimer. As a result, the authors determined 6 serum biomarkers found in the men and women that could be grouped into clusters differentiating cardiopulmonary structural and functional abnormalities (Scherzer et al., 2018). In the Allam & Gumpeny (2012) study, hierarchical clustering was used to analyze microarray data compromised of 26,097 genes and researchers found 3 genes, SORL1, APP, and APOE, highly expressed in patients diagnosed with Alzheimer's along with 21 additional genes that could potentially be linked as well. Moore et al. (2010) performed Ward's hierarchical cluster analysis identified 5 clusters with different asthma phenotypes from 34 variables and 726 participants in The Severe Asthma Research Program cohort. Each cluster was categorized as a different severity level of asthma and 80% of the subjects were assigned to the correct cluster of asthma severity (Moore et al., 2010). These studies demonstrate the power of clustering analysis as an effective tool that can be used to identify population subgroups that have specific phenotypes and properties.

Even though clustering algorithms may be sufficient to complete certain projects, sometimes they may only be needed in intermediate stages. Semi-supervised algorithms help in situations where the small portion of labeled data may be insufficient for a supervised model or when researchers are simply interested in discovering possible variables linked to their outcome variable. Bair et al. (2013) discussed the various situations where hybrid clustering models may be implemented to optimize statistical analysis and explained how clustering algorithms may still be adequate when combined with supervised algorithms, and in some cases more advantageous, in data sets were there are partial labels, prior knowledge about correlations between variables, or the outcome variable is known. For instance, in Zhu et al. (2019), a logistic regression model was created and enhanced using PCA and *k*-means clustering for the analysis of the Pima Indians Diabetes dataset. PCA was applied to the original dataset to obtain the most important variables

from the dataset to improve the performance of the *k*-means algorithm. The *k*-means algorithm was applied to further clean the data and remove any outliers to pass through the logistic regression model. Applying the PCA and *k*-means algorithm increased the logistic regression performance compared to traditional methods used in previous articles and successfully predicted a diabetes diagnosis. Similarly, Villarin (2019) analyzed 15 predictor variables from 498 census tracts in the City of Ceville in Spain to determine whether any of the variables could predict water consumption. First, 498 census tracts were reduced to 4 different groups using PCA. Through Ward's hierarchical and *k*-means clustering algorithms, 9 clusters were identified within the 4 groups. Multivariate linear regression analysis was performed on each cluster to determine any significant relationships between water consumption and the predictor variables from each cluster. The author found the average cadastral value and number of inhabitants per household were the most significant variables when predicting the quantity of water consumption (Villarin, 2019). These studies are prime examples of the versatility of clustering analysis has and its beneficial impact in various topics of interest in research.

Obesity has been a subject of interest in research due to its alarmingly increasing rates in the population (Wang et al, 2008). The magnitude of the issue has been discussed and projected that by 2030, 51.1% of adults will be considered obese (Wang et al, 2008). Studies have brought awareness to the serious side effects of obesity such as heart disease, diabetes, hypertension, and premature death in obese adults (Poirier et al, 2006; Hossain et al, 2007; Masters et al, 2013). It appears to be a linkage between obesity and inflammation based on various articles (Hotamisligil, 2006; Amin et al, 2019). The coupling of inflammation and excess weight in obese individuals may lead to further health complications (Lumeng & Saltiel, 2011). Figure 2 depicts the difference between lean and obese adipose tissue, indicating obese adipose tissue being in a pro-inflammatory

state. This has forced scientists to reassess the cause of obesity and its relationship with possible inflammatory biomarkers.

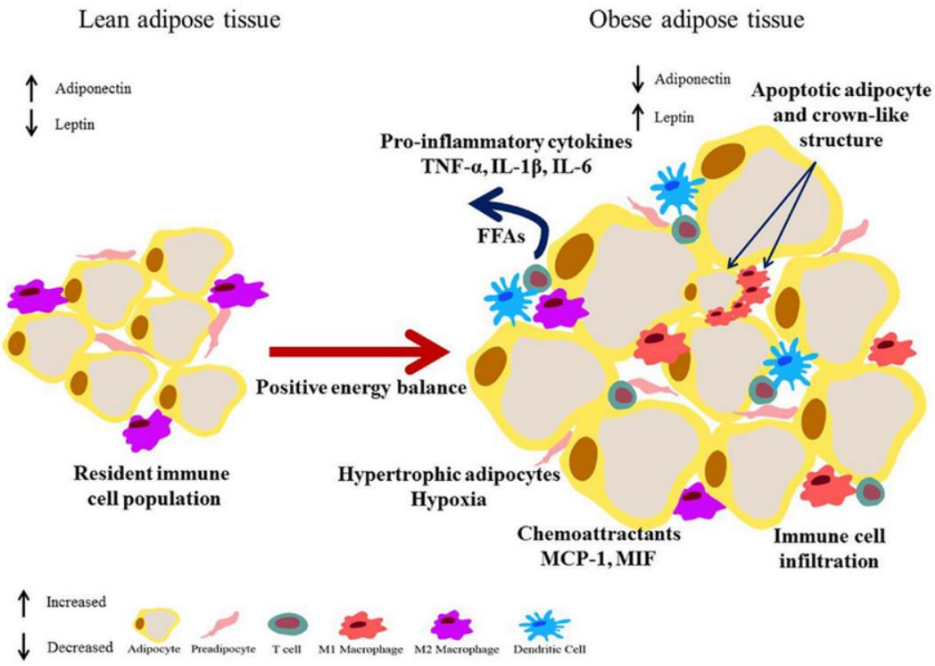

Figure 2 Difference between lean and obese adipose tissue (McArdle et al, 2013)

A few inflammatory biomarkers of interest in relation to obesity have been C-reactive protein (CRP), MCP-1, tumor necrosis factor alpha (TNFα), interleukin-6 (IL-6), and interleukin-8 (IL-8), and leptin (Wang & Nakayama, 2010; Kanda et al, 2006; Park et al, 2005). A study that included 100 patients and half were classified as obese, higher levels of MCP-1 and IL-8 were higher in patients classified as obese (Kim et al, 2006). Studies investigating the relationship between CRP and adiposity have demonstrated overweight or obese individuals tend to have higher CRP levels (Santos et al, 2005; Visser et al, 2001). Ellulu et al. (2017) proposed a mechanism of action where obesity has excess macronutrients that stimulate adipose cells to release inflammatory biomarkers such as TNFα and IL-6. Dornbush & Aeddula (2022) further discuss the similar structure between IL-6, a proinflammatory biomarker, and leptin. The authors noted the amount of leptin in the body positively correlates to the total amount of adipose tissue.

These are a few of the inflammatory biomarkers currently associated with obesity and inflammation.

Although extensive research has been carried out in this subject, the treatments for obesity remain inadequate and have led to disappointing results for patients (Brownell, 2010). It has been shown early onset of obesity will significantly increase the risk of obesity in adulthood (Serdula et al, 1993). Since the cause of obesity remains unsolved and treatment success rates remain low for adults, it's crucial to initialize early prevention efforts in the younger population. The association between obesity and inflammation biomarkers in adolescents has not been concluded and requires more research.

Clustering has advanced the research in the medical field; however, it is important to understand that some clustering methods may only be useful for specific data types. Researchers have compared different clustering methods to determine the strengths and weaknesses of each algorithm. In Hammouda & Karray (2000), *k*-means, fuzzy *c*-means, Mountain, and Subtractive clustering methods were compared in a real-life study to analyze medical data from 300 individuals and 13 different variables relating to heart disease. Each algorithm was evaluated based on their performance in correctly identifying whether an individual would ultimately be clustered with the heart disease diagnosis or not. The researchers reported that the *k*-means and fuzzy *c*-means were the most accurate algorithms, although, fuzzy *c*-means was slower than the *k*-means clustering technique. The authors decided to include mountain and subtractive clustering and discovered the data set being used in this study was too big for these algorithms and performed poorly. Yim & Ramdeen (2015) discussed the possibility of having different results while performing Hierarchical clustering analysis with 3 different linkage measures, single, average, and complete, on a psychological dataset consisting of 67 Cantonese-English bilingual young adults to try to identify

subgroups among the bilingual individuals. Although the same dataset was used for each linkage measure, different clusters were produced. The authors discussed the responsibility researchers have when choosing the right algorithm for their data and the ambiguity behind these choices. They further explain since clustering algorithms will cluster the data regardless, it does not always mean the clusters are meaningful. The interpretation and significance of the clusters are given by the researcher and if they don't have an accurate understanding of the correct clustering procedure for each algorithm, it may lead to inaccurate results (Yim & Ramdeen, 2015). These studies demonstrate the importance of understanding the concept for various clustering methods and the data being analyzed to avoid misinterpreting or influencing the results by accident.

Objectives

The objectives of our study are: (1) to compare various clustering methods in identifying inflammatory groups (biomarker phenotypes), and (2) to identify relationship between these groups with obesity measures in local adolescent population.

CHAPTER II

THEORETICAL BACKGROUND

Clustering analysis has become a powerful tool for data exploration and analysis in various fields, including machine learning, biology, and business. Jain (2010) summarizes the three different purposes for the use of cluster analysis in research as “underlying structure”, “natural classification”, and “compression.” In “underlying structure”, the researcher intends to understand the data by identifying inconsistencies and generate hypotheses. In “natural classification”, the researcher intends to establish the similarity between objects within a group and dissimilarity to the other groups. In “compression”, the goal is to summarize and condense the data into the most meaningful groups. Once the goal of the clustering analysis has been established, the researcher can determine which clustering technique is most fitting.

Clustering Algorithms

The clustering algorithms can be further categorized into two main types: hierarchical and partitional (non-hierarchical). Hierarchical algorithms merge or divide the initial clusters to build a hierarchy represented by a dendrogram. Alternatively, partitioning algorithms divide the data points into smaller partitions, or clusters, using different standards unique to each algorithm.

Furthermore, distance metrics are generally used to determine the similarity between data points in a cluster. Although there are various distance metrics available, applying an effective distance metric can greatly improve the performance of the clustering algorithm. When determining which distance metric to use, it is important to consider the data in which the

clustering analysis will be applied to. The Minkowski distance metric is the general distance formula and defined as

$$d_{mink}(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

and it can be manipulated by changing the value of p. When $p = 1$, we have the Manhattan distance defined as

$$d_{man}(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

When $p = 2$, we have the Euclidean distance defined as

$$d_{euc}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

When $p = \infty$, we have the Chebychev distance defined as

$$d_{cheb}(x, y) = \min_{i=1} |x_i - y_j|$$

For the purpose of the study, Euclidean distance was the most appropriate and used for the k-Means, Hierarchical, and Fuzzy C-Means clustering algorithms.

K-Means

The k-means clustering algorithm was first presented by MacQueen in 1967 with the main goal of grouping n data points into k clusters. Each observation or data point is matched to the cluster with the nearest centroid. The mean of the data points within the cluster is another way to define the cluster centroid. Lloyd (1982) enhanced MacQueen's k-means algorithm by incorporating "Expectation-Maximization" algorithm (Dempster et al., 1977). In Sinaga and Yang (2020), the updated k-means algorithm is thoroughly discussed as a partitioning clustering

algorithm used to partition n data points into k clusters while applying the EM algorithm. The number of clusters, k, is not known prior to the analysis and determines the number of centroids the algorithm will initiate at the start. The data points are assigned to the cluster with the nearest centroid. The centroid is defined as the mean of the cluster. The goal of the algorithm is to optimize the objective function defined as

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2$$

which is the sum of squares of distances of each object to its assigned vector, where $n = 1, .., N$ and $k = 1, \dots, K$. If $r_{nk} = 0$, the object is not assigned to the cluster. If $r_{nk} = 1$, the object is assigned to the cluster. The algorithm finds the values for r_{nk} and μ_k that minimize J. Cluster assignment and centers are updated through an iterative process, by

$$\mu_k = \frac{\sum_{n=1}^{N} r_{nk} x_{nj}}{\sum_{n=1}^{N} r_{nk}}$$

and

$$r_{nk} = \begin{cases} 1 & \text{if } \|x_n - \mu_k\|^2 = \min_{1 \le k \le K} \|x_n - \mu_k\|^2 \\ 0, & \text{otherwise} \end{cases}$$

where $\|x_n - \mu_k\|^2$ represents the Euclidean distance between the object, x_n, and the centroid, μ_k. The process will continue until the parameters have no significant change or stop changing overall, also signifying convergence is reached. The k-means algorithm assumes the derived clusters are spherical shaped and of similar size. Since the algorithm uses a distance metric for similarity between the data point and the designated centroid, the data points form a spherical shape around the centroid.

Hierarchical

Kaufman & Rousseeuw (1990) developed a method of hierarchical clustering analysis that identifies nested clusters and arranges them into a hierarchical tree, also referred to as a dendrogram. The two primary methods of hierarchical clustering are divisive analysis (DIANA) and agglomerative nesting (AGNES). Agglomerative hierarchical clustering operates using a bottom-up technique where each object begins as an individual cluster at the bottom, and as the tree ascends, similar objects combine to form new clusters at the top. On the other hand, divisive hierarchical clustering, follows a top-down approach, where the data begins as one cluster, and as the tree descends, the initial cluster breaks into new clusters to combine the most similar data points. This process ultimately leads to smaller clusters as it reaches the bottom. The degree of similarity between the objects is computed using a distance formula. There exist several distance formulas that can be applied in hierarchical clustering analysis, but the Euclidean distance is often the default option for clustering algorithms and was used in this analysis.

Additionally, hierarchical clustering involves four primary linkage methods: single, average, complete, and Ward's method. Ward's method was proposed by Ward (1963) as a linkage criterion for hierarchical clustering. When clusters are joined or divided, the aim is to reduce the variation within a cluster as much as possible. The procedure estimates the total of the squared distances between each data point in the cluster and the centroid in order to determine the variance of the produced clusters.

Given C, a group of data points, the error sum of squares (ESS) associated with C is calculated by

$$\mathrm{ESS}(C) = \sum_{x \in C} \left(\mathbf{x} - \mu(C)\right)\left(\mathbf{x} - \mu(C)\right)^T = \sum_{x \epsilon C} \mathbf{x}\mathbf{x}^T - |C|\mu(C)\mu(C)^T$$

where $\mu(C)$ is the mean of C,

$$\mu(C) = \frac{1}{|C|} \sum_{x \epsilon C} \mathbf{x}$$

Assuming there are k groups in one level of the clustering, $C_1, C_2, \ldots, C_k$, then the amount of information that is lost is expressed as the sum of the error sum of squares, which is calculated by

$$\text{ESS} = \sum_{i=1}^{k} \text{ESS}\ (C_i)$$

which is the total within-group ESS.

Fuzzy C-Means

Bezdek (1984) presented the Fuzzy C-means method as an enhancement to Lloyd's (1982) classic *k*-means algorithm. The fundamental distinction between k-means and fuzzy *c*-means clustering algorithms is their methodologies for assigning each data point to a cluster. Bezdek's contribution was to allow for a degree of membership of each data point in each cluster, as opposed to the hard clustering of *k*-means, which assigns each data point to just one cluster based on its proximity to the centroid. Another main difference is in the algorithms' assumptions. The FCM algorithm does not assume clusters will be spherical or that all clusters will be equal in size, as assumed in k-means. The FCM algorithm aims to minimize the objective function defined as

$$J_m = \sum_{i=1}^{D} \sum_{j=1}^{N} \mu_{ij}^{m} \left\| x_i - c_j \right\|^2$$

where D is the number of data points and N denotes the number of clusters. The value $m > 1$ in the fuzzy partition matrix is applied as an exponent that determines the extent of fuzzy overlap

across the clusters. The number of data points that have significant membership in more than one cluster is used to calculate the overlap. The degree of fuzziness in the membership assignments is influenced by the value of m, where x_i represents the i^{th} data point, c_j denotes the center of the j^{th} cluster, and μ_{ij} represents the extent to which x_i belongs to the j^{th} cluster. Consequently, the form and size of the resulting clusters are affected by the value of m.

In the FCM algorithm, each data point, x_i is assigned membership values. These values are initialized randomly at the start. It should be emphasized that for any data point, the total membership values across all clusters add up to 1, which reflects the assumption that each point belongs to some degree to each of the resulting clusters. The cluster centers are calculated by

$$c_j = \frac{\sum_{i=1}^{D} \mu_{ij}^{m} \, x_i}{\sum_{i=1}^{D} \mu_{ij}^{m}}$$

And updating the μ_{ij} by

$$\mu_{ij} = \frac{1}{\sum_{i=1}^{N} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

Finally, the algorithm will calculate the objective function, J_m, while repeating the previous steps until convergence is reached. Convergence can be attained by achieving a certain minimum threshold improvement value or by performing a specified maximum number of iterations to enhance the J_m. Improving the J_m by less than a specified minimum threshold, or until a specified maximum number of iterations, will allow for convergence to be reached.

Gaussian Mixture Model

The Gaussian Mixture Model (GMM) is a clustering algorithm proposed by Dempster et al. (1977) that assumes the data arise from a mixture of multiple Gaussian distributions, where

each is associated with a different cluster. The algorithm employs the Expectation-Maximization (EM) method and probabilistic models to estimate the parameters of the Gaussian mixture and assign data points to clusters based on their probabilities. Specifically, the EM algorithm iteratively refines the estimates of the mixture parameters and the posterior probabilities of each data point belonging to each cluster until convergence.

To cluster the data using the GMM algorithm, the parameters and the probabilities of each data point belonging to each cluster are initially assigned random values, then updated iteratively until convergence is achieved. Once it converges, the data points are assigned to the most likely cluster based on the probabilities computed in the "Expectation" (E) step.

The GMM objective function is to maximize the likelihood value for data X, the likelihood value, $p(x)$,

$$p(x_i) = \sum_{k=1}^{K} p(x_i|c_k)\, p(c_k)$$

$$p(X) = \prod_{i=1}^{N} p(x_i) = \prod_{i=1}^{N} \sum_{k=1}^{K} p(x_i|c_k)\, p(c_k)$$

or the log-likelihood value, L,

$$L = \log\big(p(X)\big) = \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} p(x_i|c_k)\, p(c_k)\right)$$

If we assume that the data were generated by a mixture of K Gaussian distributions, we can express the probability density function, $p(x)$ as the sum of the marginalized probabilities of each cluster for all data points. To estimate the maximum likelihood parameters of the Gaussian mixture, we can apply the Expectation Maximization (EM) algorithm, which is which is often

used in practice and yields the maximum likelihood estimates. The EM algorithm consists of computing the probability, r_{ic}, that the data point, x_i, belong to cluster c through

$$r_{ic} = p(y_c \mid x_i) = \frac{p(y_c, x_i)}{p(x_i)} = \frac{p(x_i \mid y_c)\ p(y_c)}{\sum_{c=1}^{K} p(x_i|y_c)\ p(y_c)}$$

After computing the probabilities of each data point belonging to each cluster, the next step in the algorithm is to determine the new parameter m_c, which specifies the proportion of points assigned to each cluster. The parameters are updated by calculating the maximum likelihood estimates for each cluster c through

$$m_c = \sum_{i=1}^{N} r_{ic}\ ,$$

$$\pi_c = \frac{m_c}{N}\ ,$$

$$\mu_c = \frac{1}{m_c} \sum_{i=1}^{N} r_{ic}\ x_i\ ,$$

$$\sum\nolimits_c = \frac{1}{m_c} \sum_{i=1}^{N} r_{ic}\ (x_i - \mu_c)^T\ (x_i - \mu_c)$$

To estimate the unknown parameters of a mixture of Gaussian distributions, the Expectation-Maximization (EM) algorithm is iteratively applied until the log-likelihood value, L, converges. The GMM method assumes the observed data set was generated from a mixture of Gaussian (normal) distributions with unknown parameters. Each cluster is modeled as a normal distribution characterized by its mean and covariance.

Principal Component Analysis

Introduced by Pearson (1901), Principal Component Analysis (PCA) is a statistical method that aims to reduce the dimensionality of a dataset. Its objective is to identify the

principal components, which are linear combinations of the dataset features that capture the highest variance in the data. These principal components are obtained as the eigenvectors of the covariance matrix of the data, and the corresponding eigenvalues represent the amount of variance explained by each component. The principal components are ranked in order of the eigenvalues, and the first component explains the most variance in the data. All the principal components combined explain the total variance of the dataset. Although the number of principal components is equal to the number of variables in the dataset, the first component accounts for the largest variance overall, while the subsequent components account for decreasing amounts of variance and are perpendicular to the preceding ones. It is important to note that all principal components are calculated under the condition of being orthogonally rotated, or perpendicular, to avoid correlation with each other, thus explaining the maximum amount of variance in the data set.

Before initiating the analysis, the data must be standardized by subtracting the mean and dividing by the standard deviation for each variable. Once the variables have been standardized, the covariance matrix must be computed. The matrix is a d x d matrix, where d is the number of variables, thus demonstrating symmetry. The entries in the covariance matrix represent the covariance between each pair of variables.

Once the principal components are obtained, the data is projected on a lower-dimensional space made up by the first k principal components, where k is the number of dimensions of the new subspace.

CHAPTER III

METHODS

Study Data

This study was conducted using secondary data collected from 183 local adolescents. Data included socio-demographic and anthropometric measures, such as age, sex, height, weight, head circumference, body mass index (BMI), body circumferences to assess for adiposity (waist and hip), along with 6 different inflammatory biomarkers: CRP, HGF, IL-8, Leptin, TNFα, and MCP1. The study focused only on the 6 inflammatory biomarkers and their association with BMI and waist circumference.

Data Analysis

Missing values were removed and a total of 167 observations with complete data were analyzed. All the variables, except MCP1, had a right-skew distribution and where log transformed to obtain a normal, or approximately normal, distribution. Normality was evaluated using histograms, normal probability plots, and Shapiro-Wilk test. All variables were standardized for the purpose of the clustering analysis.

Descriptive statistics of the study population were produced for the variables considered in the analyses. Continuous variables were described with mean and standard deviation and categorical variables were described with frequencies and percentages. Spearman correlation analysis was conducted to determine whether any variables were correlated with each other.

Five different clustering algorithms were applied to the data: *k*-Means, Ward's hierarchical agglomerative method, fuzzy *c*-means, Gaussian mixture model, and principal component analysis. The clustering algorithms were computed using R software, while mainly using the "cluster" and "factoextra" package (Maechler et al, 2013; Kassambara & Mundt, 2017). The "cluster" package was used for the computation of *k*-Means, Ward's hierarchical agglomerative method, and fuzzy *c*-means. The "mclust" package was used for the computation of the Gaussian mixture model clusters. The "factoextra" package was used for the graphical representations of the derived clusters and for the statistical methods needed for validating the number of clusters used for the *k*-Means and Ward's hierarchical agglomerative method. The three methods included in this package and were used were the elbow, silhouette, and gap statistic. These statistics were not applicable to the fuzzy *c*-means or Gaussian mixture model.

The "clValid" and "fpc" package were used to obtain the cluster validation indices, including: Dunn Index, Average Silhouette width, Calinski-Harabasz index, and the within-cluster sum of squares (Brock et al, 2008; Hennig & Imports, 2015). A maximized Dunn index indicates a better clustering outcome. The average silhouette width ranges from 0 to 1, where a value closer to 1 suggests the data are better clustered. A higher Calinski-Harabasz index indicates better clustering performance. The within-cluster sum of squares evaluated internal cohesion and external separation. A smaller value indicated more closely related objects within the cluster.

Multinomial logistic regression was used to determine factors associated with cluster membership, as well as to evaluate the relationship of the derived clusters and BMI and waist circumference. The principal component analysis was conducted using SAS 9.4 (SAS Institute,

Inc). Tukey-Kramer test was used for post-hoc multiple pairwise comparisons of means between groups. All statistical tests were two-sided and were performed at significance level of 0.05.

CHAPTER IV

RESULTS

Statistical Analysis

Descriptive Statistics

The study population was composed of the local adolescent population whose age ranged from 14-20 years old. The majority of the study population consisted of sex1, 63.4%, and BMI ranging from 17.50 - 49.36 (Table 1). The normality of the distribution of each of the variables was evaluated with histograms, normal probability plots and the Shapiro-Wilk test. Variables with right skewed distribution, such as CRP, HGF, IL-8, Leptin, and TNFα, were log-transformed for normalization and be used in the clustering analysis.

Table 1 Descriptive statistics of study data

Numerical	
Variable	**Mean (SD)**
Age	15.79 (1.40)
BMI	25.76 (6.42)
Waist Circumference	86.00 (17.32)
CRP	1944.57 (3836.57)
HGF	1166.87 (821.87)
IL-8	2.31 (1.38)

Table 1, continued

Leptin	15062.39 (16503.55)
TNFα	3.25 (1.24)
MCP1	218.37 (104.34)
lCRP	6.71 (1.16)
lHGF	6.78 (0.88)
lIL-8	0.74 (0.41)
lLeptin	9.16 (1.05)
lTNFα	1.13 (0.36)
Categorical	
Variable	**Frequency**
Sex 0	67 (36.6%)
Sex 1	116 (63.4%)

Figure 3 is a heat map created to represent the Spearman correlation coefficients, r, between the 6 biomarkers used in the clustering analysis. The red color represents a negative correlation, and the blue color represents a positive correlation. The intensity of the color will depend on the strength of the correlation. Most, if not all, relationships between the variables are none or very weak ($r < .3$). There appears to be a weak relationship between CRP and Leptin ($r=0.41$) and IL-8 and TNFα ($r=0.37$).

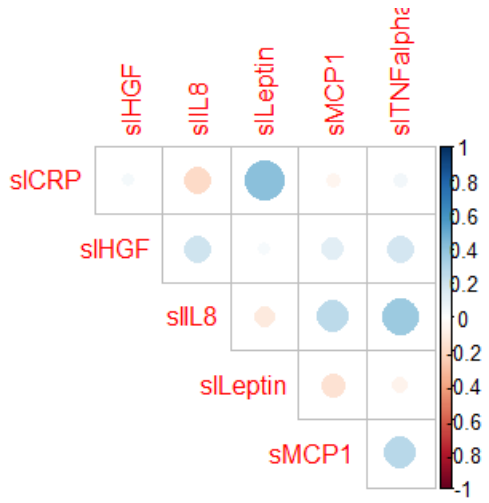

Figure 3 Spearman correlation heat map

Cluster Analysis

***K*-Means.** *K*-Means cluster analysis was performed on standardized log-transformed biomarkers. The optimal number of clusters was determined by the elbow, silhouette, and gap statistic methods (Figure 4), as well as visual inspection of the derived clusters produced by the algorithm (Figure 5, Figure 6). The gap statistic method did not produce an optimal number of clusters, while the elbow and silhouette method suggested that 2 or 3 clusters were optimal. *K*-means clusters of k=2 and k=3 produced the best cluster separation, meaning the overlap between clusters was minimized (Figure 5, Figure 6).

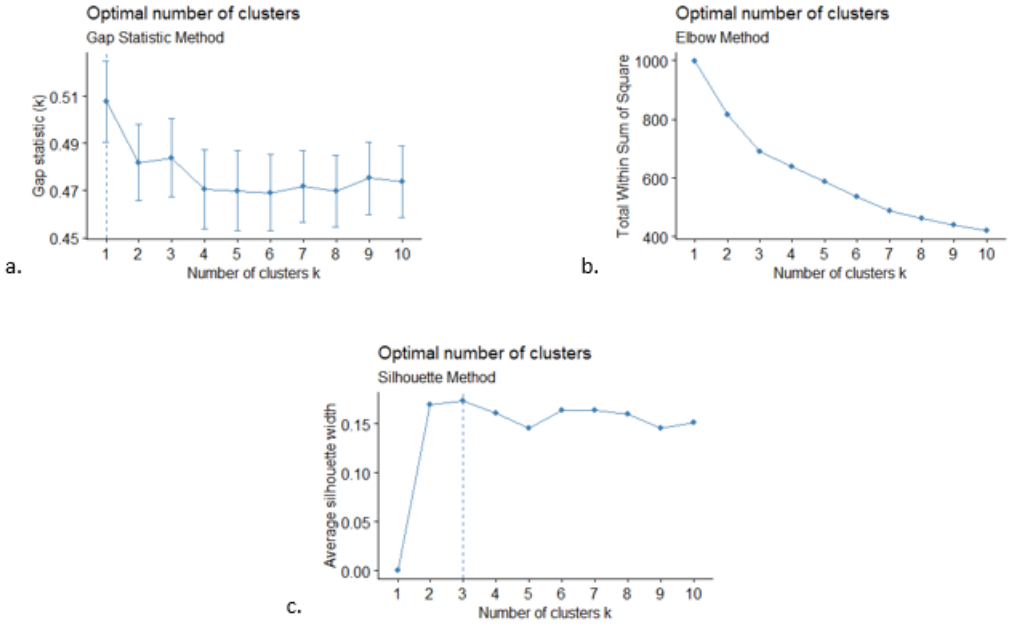

Figure 4 Optimal number of clusters for *K*-Means algorithm

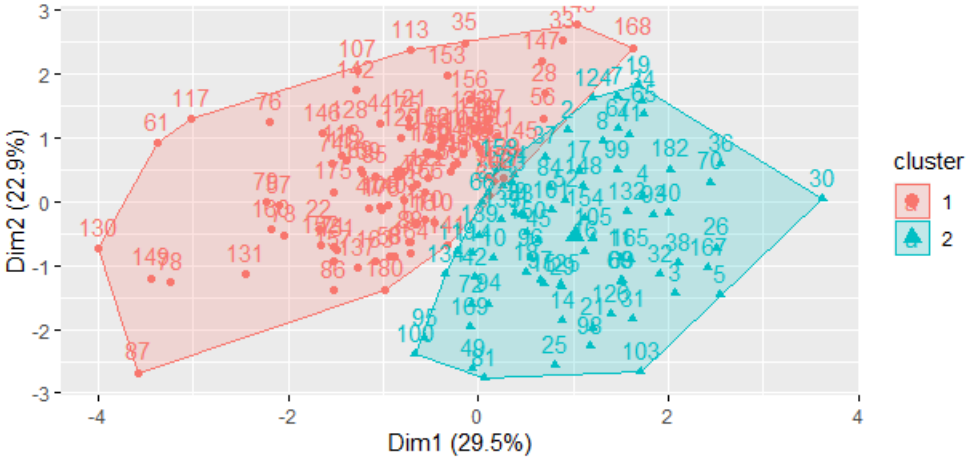

Figure 5 Clusters from *K*-Means algorithm (k=2)

Table 2 Clusters from *K*-Means algorithm (k=2)

Biomarker	Cluster 1 (n=92)	Cluster 2 (n=75)	p-value
lCRP	6.14 (0.75)	7.40 (1.20)	<0.0001
lHGF	6.83 (0.89)	6.72 (0.87)	0.4371
lIL-8	0.91 (0.43)	0.54 (0.28)	<0.0001
lLeptin	8.65 (0.95)	9.79 (0.80)	<0.0001
lTNFα	1.22 (0.35)	1.01 (0.33)	<0.0001
MCP1	261.92 (105.27)	164.95 (74.47)	<0.0001

Table 2 shows the descriptive characteristics of the derived clusters for k-means (k=2). A t-test used to compare the means between two groups, showed that Cluster 1 compared to Cluster 2 was characterized with higher mean levels of ln(IL-8) ($p<0.0001$), ln(TNF- α) ($p<0.0001$), and MCP1 ($p<0.0001$). Cluster 2 was distinguished from Cluster 1 with higher mean levels of ln(CRP) (<0.0001) and ln(Leptin) ($p<0.0001$).

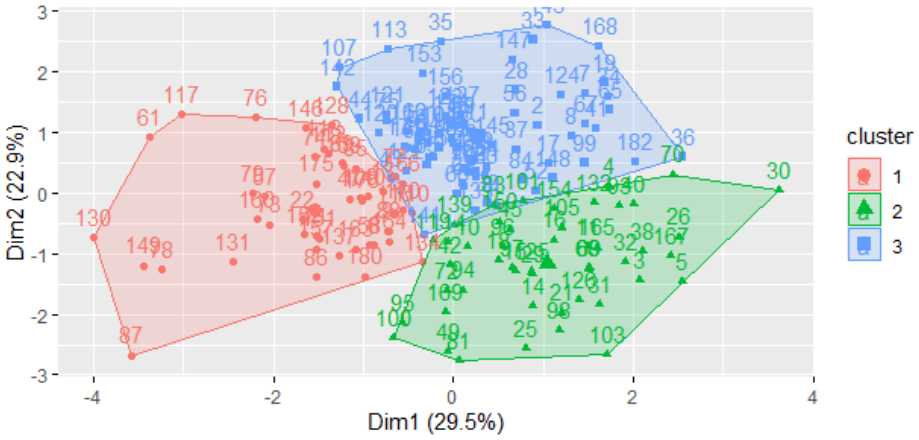

Figure 6 Clusters from *K*-Means algorithm (k=3)

Table 3 Clusters from *K*-Means algorithm (*k*=3)

Biomarker	Cluster 1 (n=46)	Cluster 2 (n=73)	Cluster 3 (n=48)	p-value
lCRP	6.44 (0.88)^	6.03 (0.60)^	8.00 (0.97)^	<0.0001
lHGF	7.18 (0.61)^	6.54 (0.96)^	6.77 (0.86)	0.0005
lIL8	1.14 (0.45)^*	0.63 (0.29)^	0.54 (0.26)*	<0.0001
lLeptin	8.95 (0.98)^	8.69 (0.92)*	10.10 (0.64)^*	<0.0001
lTNFα	1.46 (0.25)^	0.94 (0.27)^	1.09 (0.34)^	<0.0001
MCP1	287.63 (110.58)^*	204.72 (93.10)^	172.75 (78.12)*	<0.0001

^* indicate difference in means as determined by Tukey-Kramer test.

Table 3 shows the descriptive characteristics of the derived clusters for *k*-means (k=3). Tukey-Kramer test for post-hoc multiple pairwise comparisons of means between groups, showed that Cluster 1 compared to Cluster 2 and Cluster 3 was characterized with higher mean levels of ln(HGF) ($p=0.0005$), ln(IL-8) ($p<0.0001$), ln(TNF- α) ($p<0.0001$), and MCP1 ($p<0.0001$). Cluster 3 compared to Cluster 1 and Cluster 2 was characterized with higher mean levels of ln(CRP) ($p<0.0001$) and ln(Leptin) ($p<0.0001$). Cluster 2 compared to Cluster 1 was characterized with lower mean levels of ln(CRP) ($p<0.0001$), ln(HGF) ($p=0.0005$), ln(IL-8) ($p<0.0001$), ln(TNFα) ($p<0.0001$), and MCP1 ($p<0.0001$). Cluster 2 compared to Cluster 3 was characterized with lower mean levels of ln(CRP) ($p<0.0001$), ln(Leptin) ($p<0.0001$), and ln(TNFα) ($p<0.0001$).

Hierarchical (Ward's). Hierarchical cluster analysis was performed on standardized log-transformed biomarkers. The optimal number of clusters was determined by the elbow, silhouette, and gap statistic methods (Figure 7), as well as visual inspection of the derived clusters produced by the algorithm (Figure 8, Figure 9). The gap statistic method did not produce an optimal number of clusters, while the elbow and silhouette method suggested 2 or 6 clusters were optimal. Based on the plots of the derived clusters with *k*=2 and *k*=3, hierarchical clusters of *k*=2 and *k*=3 produced good cluster separation of the analyzed data (Figure 8, Figure 9).

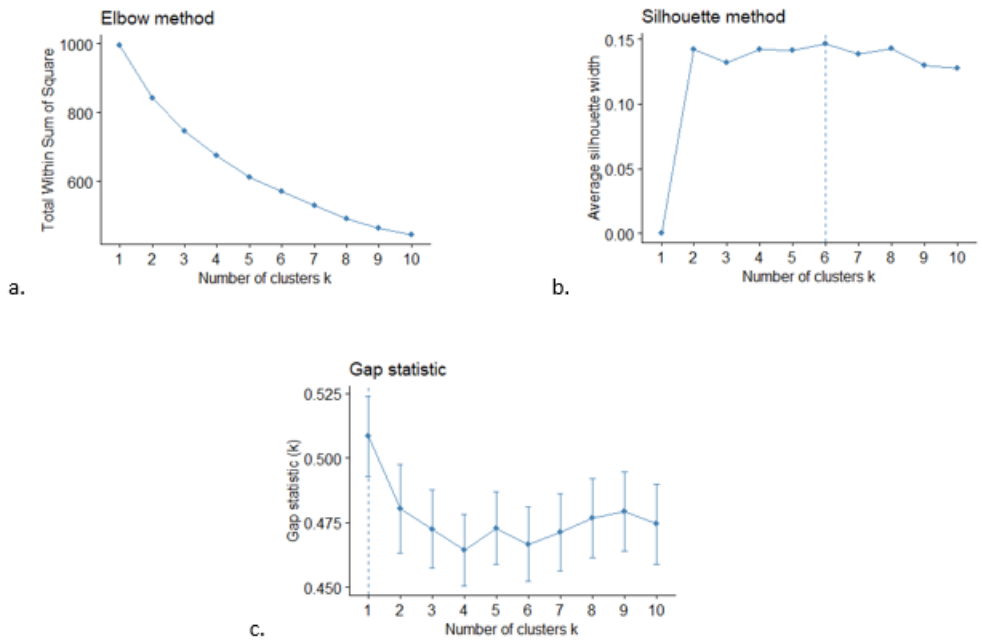

Figure 7 Optimal number of clusters for Hierarchical (Ward's) algorithm

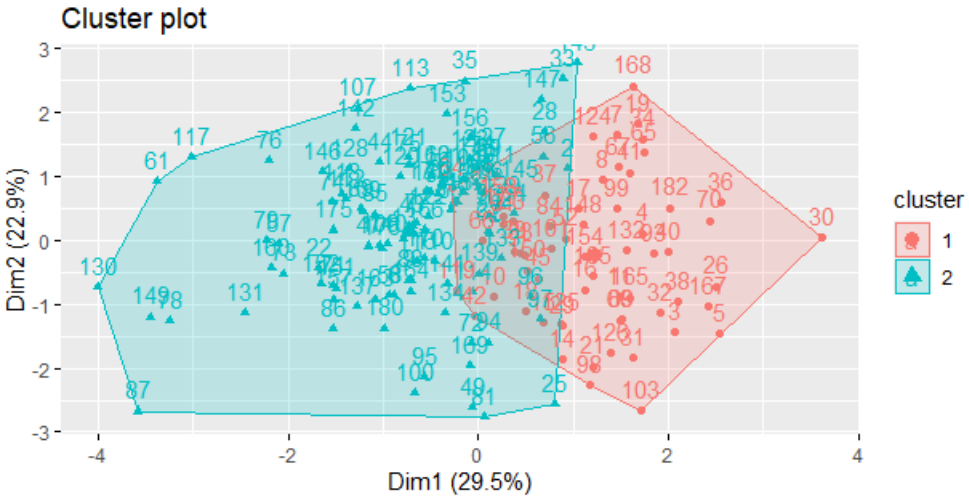

Figure 8 Clusters from Hierarchical algorithm (k=2)

Table 4 Clusters from Hierarchical algorithm (k=2)

Biomarker	Cluster 1 (n=117)	Cluster 2 (n=50)	p-value
lCRP	6.45 (1.09)	7.32 (1.09)	<0.0001
lHGF	6.76 (0.98)	6.83 (0.61)	0.0083
lIL-8	0.86 (0.41)	0.47 (0.26)	<0.0001
lLeptin	8.86 (1.00)	9.88 (0.79)	<0.0001
lTNFα	1.22 (0.36)	0.90 (0.25)	<0.0001
MCP1	245.57 (107.57)	154.73 (59.93)	<0.0001

Table 4 shows the descriptive characteristics of the derived clusters for Hierarchical (k=2). A t-test used to compare the means between two groups, showed that Cluster 2 compared to Cluster 1 was characterized with higher mean levels of ln(CRP) (<0.0001) and ln(Leptin) (p<0.0001). Cluster 1 was distinguished from Cluster 2 with higher mean levels of ln(HGF) (p=0.0083), ln(IL-8) (p<0.0001), ln(TNF- α) (p<0.0001), and MCP1 (p<0.0001).

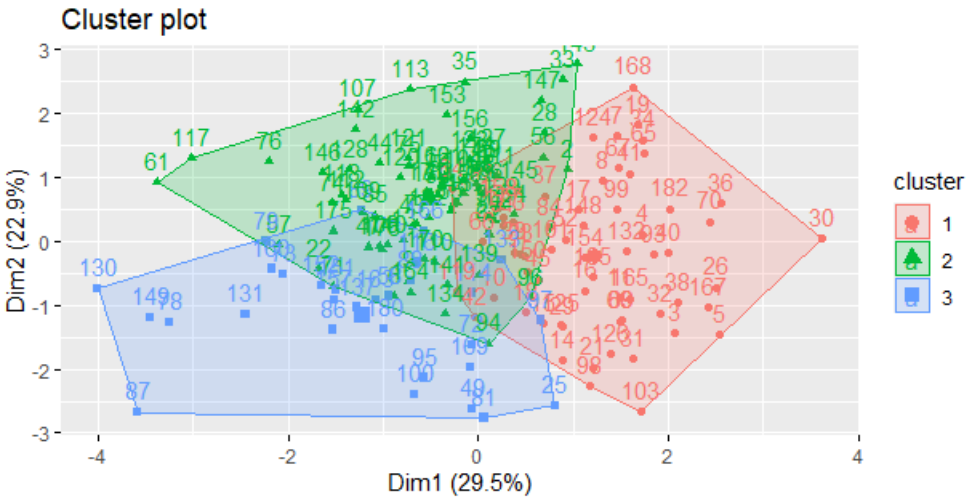

Figure 9 Clusters from Hierarchical algorithm (k=3)

Table 5 Descriptive characteristics of the clusters from Hierarchical algorithm (k=3)

Biomarker	Cluster 1 (n=28)	Cluster 2 (n=50)	Cluster 3 (n=89)	p-value
lCRP	7.61 (1.26)^	7.32 (1.10)*	6.08 (0.71)^*	<0.0001
lHGF	7.20 (0.66)^	6.83 (0.61)	6.62 (1.02)^	0.0083
lIL8	0.96 (0.45)^	0.47 (0.26)^*	0.83 (0.40)*	<0.0001
lLeptin	9.31 (0.82)^	9.88 (0.79)^	8.72 (1.01)^	<0.0001
lTNFα	1.51 (0.29)^	0.90 (0.25)^	1.13 (0.33)^	<0.0001
MCP1	333.42 (88.94)^	154.73 (59.93)^	217.93 (97.97)^	<0.0001

^* indicate difference in means as determined by Tukey-Kramer test.

Table 5 shows the descriptive characteristics of the derived clusters for Hierarchical (k=3). Tukey-Kramer test for post-hoc multiple pairwise comparisons of means between groups, showed that Cluster 2 compared to Cluster 1 and 3 had higher mean levels of ln(Leptin) ($p<0.0001$). Cluster 1 compared to Cluster 2 and Cluster 3 was characterized with higher mean levels of ln(TNF- α) ($p<0.0001$), and MCP1 ($p<0.0001$). Cluster 3 compared to Cluster 1 was characterized with lower mean levels of ln(CRP) ($p<0.0001$), ln(HGF) ($p=0.0005$), ln(TNFα) ($p<0.0001$), and MCP1 ($p<0.0001$). Cluster 3 compared to Cluster 2 was characterized with lower mean levels of ln(CRP) ($p<0.0001$), ln(Leptin) ($p<0.0001$), ln(TNFα) ($p<0.0001$), and MCP1 ($p<0.0001$).

Fuzzy *C*-Means. Fuzzy c-means cluster analysis was performed on the 6 standardized and log-transformed biomarkers. Fuzzy clusters of c=2 and c=3 produced the best cluster plots as seen in Figure 10 and Figure 11.

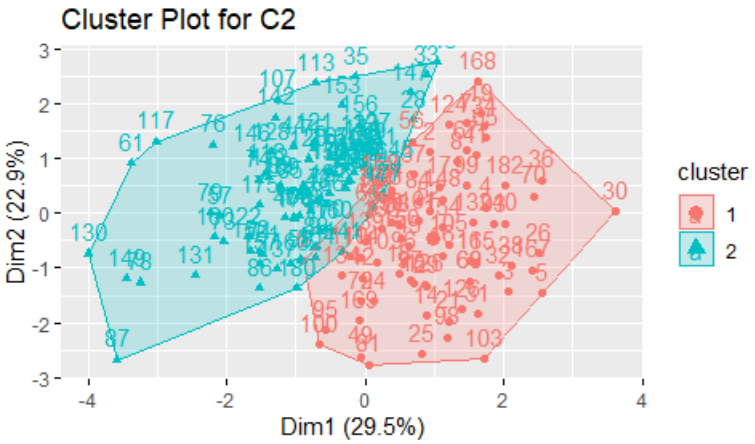

Figure 10 Clusters from Fuzzy C-Means algorithm (c=2)

Table 6 Clusters from Fuzzy C-Means algorithm (c=2)

Biomarker	Cluster 1 (n=87)	Cluster 2 (n=80)	p-value
lCRP	6.45 (1.03)	7.16 (1.19)	<0.0001
lHGF	6.83 (0.92)	6.72 (0.87)	0.4602
lIL-8	0.79 (0.42)	0.63 (0.37)	0.0090
lLeptin	8.91 (0.99)	9.60 (0.85)	<0.0001
lTNFα	1.12 (0.33)	1.09 (0.38)	0.5860
MCP1	232.07 (111.05)	198.10 (96.82)	0.0394

Table 6 shows the descriptive characteristics of the derived clusters for Fuzzy c-means (c=2). A t-test used to compare the means between two groups, showed that Cluster 2 compared to Cluster 1 was characterized with higher mean levels of ln(CRP) (<0.0001) and ln(Leptin) (p<0.0001). Cluster 1 was distinguished from Cluster2 with higher mean levels of ln(IL-8) (p=0.0090) and MCP1 (p=0.0394). There was no significant difference in mean levels of ln(HGF) and ln(TNFα) between Cluster 1 and Cluster 2.

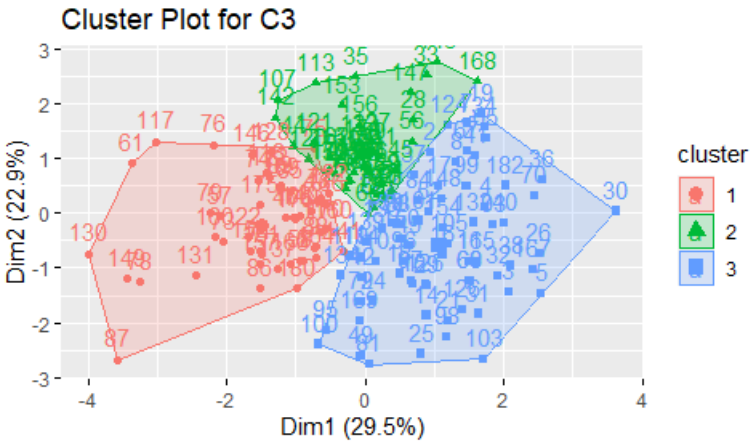

Figure 11 Clusters from Fuzzy C-Means algorithm (c=3)

Table 7 Clusters from Fuzzy C-Means algorithm (c=3)

Biomarker	Cluster 1 (n=50)	Cluster 2 (n=68)	Cluster 3 (n=49)	p-value
lCRP	6.46 (0.88)*	7.25 (1.22)^*	6.50 (1.14)^	<0.0001
lHGF	6.57 (1.03)^	6.74 (0.88)	7.04 (0.71)^	0.0393
lIL-8	0.76 (0.446)	0.63(0.39)	0.78 (0.35)	0.0781
lLeptin	8.94 (0.78)*	9.66 (0.90)^*	8.96 (1.10)^	<0.0001
lTNFα	1.10 (0.35)	1.09 (0.38)	1.13 (0.33)	0.8856
MCP1	215.70 (114.68)	198.92 (98.60)	239.31 (102.55)	0.1318

^* indicate difference in means as determined by Tukey-Kramer test.

Table 7 shows the descriptive characteristics of the derived clusters for Fuzzy c-means (c=3). Tukey-Kramer test for post-hoc multiple pairwise comparisons of means between groups, showed that Cluster 3 compared to Cluster 1 and 2 was characterized with higher mean levels of ln(HGF) (p=0.0393) and ln(IL-8) (p=0.0781). Cluster 2 compared to Cluster 1 and Cluster 3 was characterized with higher mean levels of ln(CRP) (<0.0001) and ln(Leptin) (p<0.0001). Cluster 1 compared to Cluster 3 was characterized with lower mean levels of ln(CRP) (p<0.0001) and ln(HGF) (p=0.0022). Cluster 1 compared to Cluster 2 was characterized with lower mean levels of ln(CRP) (p<0.0001) and ln(Leptin) (p<0.0001).

Gaussian Mixture Model. Gaussian Mixture Model cluster analysis was performed on standardized log-transformed biomarkers. The optimal number of clusters was calculated using the "G", the optimal number of mixture components, value from the "Mclust" function in R Studio. Figure 12 shows the 2 clusters constructed using the Gaussian Mixture Model.

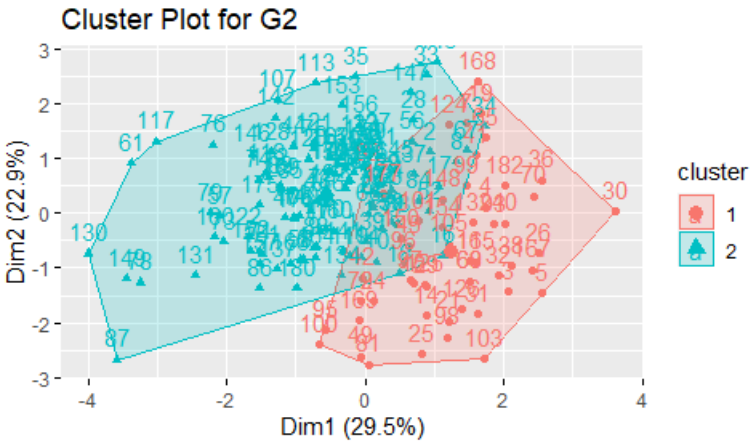

Figure 12 Clusters from Gaussian Mixture model algorithm (*g*=2)

Table 8 Clusters from Gaussian Mixture model algorithm (*g*=2)

Biomarker	Cluster 1 (n=53)	Cluster 2 (n=114)	p-value
lCRP	7.73	6.23	<0.0001
lHGF	6.42	6.95	0.0002
lIL-8	0.50	0.86	<0.0001
lLeptin	9.89	8.83	<0.0001
lTNFα	1.02	1.17	0.0100
MCP1	170	241	<0.0001

Table 8 shows the descriptive characteristics of the derived clusters for Gaussian Mixture Model (*g*=2). A t-test used to compare the means between two groups, showed that Cluster 1 compared to Cluster 2 was characterized with higher mean levels of ln(CRP) (<0.0001) and ln(Leptin) (p<0.0001). Cluster 2 was distinguished from Cluster 1 with higher mean levels of ln(HGF) (p=0.0002), ln(IL-8) (p<0.0001), ln(TNF- α) (p<0.01), and MCP1 (p<0.0001).

Validation Indices. Four different indices were calculated to compare the quality of the clusters created by the k-means, Hierarchical, Fuzzy c-means, and Gaussian Mixture Model algorithms. The calculated indices were Dunn Index, Average Silhouette width, Calinski-Harabasz index, and the within-cluster sum of squares. A maximized Dunn index indicates a better clustering outcome. Hierarchical clustering algorithm had the highest Dunn index value, 0.1122, equally for k=2 and k=3 clusters, followed by Gaussian mixture model, with a Dunn index value of 0.1054. The average silhouette width ranges from 0 to 1, where a value closer to 1 suggests the data are better clustered. Gaussian mixture model had the closest average silhouette width value equal to 0.1757, followed by k-means (k=3) with an average silhouette width equal to 0.1730. A higher Calinski-Harabasz index indicates better clustering performance. *K*-means (k=2) had the highest Calinski-Harabasz index value, 37.166, followed by Fuzzy c-means (c=2) with a Calinski-Harabasz index value of 36.738 and K-means (k=3) had a value of 36.658. The within-cluster sum of squares evaluated internal cohesion and external separation. A smaller value indicated more closely related objects within the cluster. K-means (k=3) had the lowest within-cluster sum of squares of 688.295, followed by fuzzy c-means (c=3) with a value of 718.508. A summary of all the validation indices is shown in Table 9.

Table 9 Cluster validation indices

Validation Index	K-Means		Hierarchical		Fuzzy C-Means		Gaussian
	k=2	k=3	k=2	k=3	c=2	c=3	g=2
Dunn	0.0995	0.0969	0.1122	0.1122	0.1046	0.0970	0.1054
Avg.silwidth	0.1695	0.1730	0.1425	0.1323	0.1687	0.1356	0.1757
Calinski-Harabasz	37.166	36.658	30.5368	27.3551	36.738	31.669	31.8725
Within.ss	812.898	688.295	840.456	746.851	814.622	718.508	834.754

Principal Component Analysis. Principal component analysis was performed on the 6 different biomarkers. The principal axis method was used to extract the components, and this was followed by a varimax (orthogonal) rotation. Only the first two components had eigenvalues greater than 1.00; results of a scree plot also suggested that only the first two were meaningful. On Figure 13, the eigenvalue significantly drops after factor 2, meaning the remaining factors account for a smaller amount of the total variance, therefore, only the first two components were retained for rotation.

Table 10 Rotated factor pattern and final community estimates from PCA

Component 1	Component 2	h^2	Biomarker
4	80	0.65	CRP
62	13	0.41	HGF
67	-19	0.49	IL8
-8	70	0.50	Leptin
62	-16	0.41	MCP1
69	10	0.49	TNFα

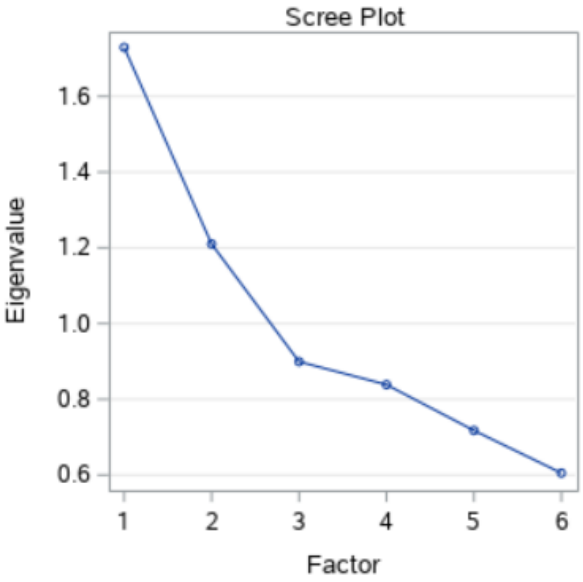

Figure 13 PCA scree plot

Regression Analysis

Multinomial Logistic. Table 11 shows the results of multinomial logistic regression analysis of factors associated with the 2 derived clusters from the k-means, Hierarchical Ward's Method, Fuzzy c-means, and Gaussian Mixture Model algorithms. Cluster 1 was characterized as the baseline, or healthy cluster, for the analysis.

For the k-Means algorithm, adolescents with higher levels of BMI and higher waist circumference measurements had significantly higher odds of membership in Cluster 1, which was characterized with higher levels of lCRP and lLeptin compared to Cluster 2 as seen in Table 2. For the Hierarchical Ward's Method, adolescents with higher levels of BMI and higher waist circumference measurements had significantly higher odds of membership in Cluster 1, which was characterized with higher levels of lCRP and lLeptin compared to Cluster 2 as seen in Table 4. For Fuzzy c-means, adolescents with higher levels of BMI and higher waist circumference measurements had significantly higher odds of membership in Cluster 1, which was characterized with higher levels of lCRP and lLeptin compared to Cluster 2 as seen in Table 6. The Gaussian Mixture Model did not produce any significant results.

Table 12 shows the results of multinomial logistic regression analysis of factors associated with the 3 derived clusters from the k-means, Hierarchical Ward's Method, and Fuzzy c-means algorithms. Cluster 3 was characterized as the baseline, or healthy cluster, for the analysis.

The k-Means algorithm demonstrated adolescents with higher levels of BMI had significantly higher odds of membership in Cluster 2, which was characterized with higher levels of lCRP and lLeptin compared to Cluster 3 as seen in Table 3. In addition, when BMI was adjusted for age and sex, higher levels of BMI had significantly higher odds of membership in Cluster 2,

which was characterized with higher levels of lCRP and lLeptin compared to Cluster 3. Adolescents with a higher waist circumference had significantly higher odds of membership in Cluster 2 compared to Cluster 3, including when adjusted for age and sex as well.

Similar results can also be seen for the Fuzzy c-means algorithm. Adolescents with higher levels of BMI and higher waist circumference measurements had significantly higher odds of membership in Cluster 1, which was characterized with higher levels of lCRP and lLeptin compared to Cluster 3 as seen in Table 7. The same results were given when adjusting for age and sex.

As for the results for Hierarchical Ward's method, the results weren't as clear as the previous algorithms. There wasn't a definite distinction of higher odds membership between the Cluster 1 or Cluster 2 compared to Cluster 3 for higher levels of BMI. Adolescents with a higher waist circumference had significantly higher odds of membership in Cluster 2 compared to Cluster 3, which was characterized with higher levels of lLeptin compared to Cluster 3 as seen in Table 5.

Table 11 Crude and Adjusted OR (95% CI) based on logistic regression for 2 derived clusters

Clustering Method	Variable	Crude		Adjusted**	
		Cluster 1 vs. Cluster 2*		Cluster 1 vs. Cluster 2*	
		OR (95% CI)	*p*-value	OR (95% CI)	*p*-value
K-Means	BMI	1.219 (1.134, 1.309)	<0.0001	1.279 (1.172, 1.397)	<0.0001
	Waist	1.057 (1.032, 1.082)	<0.0001	1.070 (1.042, 1.099)	<0.0001
Hierarchical Ward's	BMI	1.182 (1.109, 1.260)	<0.0001	1.210 (1.125, 1.301)	<0.0001
	Waist	1.052 (1.029, 1.076)	<0.0001	1.062 (1.036, 1.089)	<0.0001
Fuzzy *C*-Means	BMI	1.129 (1.067, 1.196)	<0.0001	1.169 (1.095, 1.247)	<0.0001
	Waist	1.043 (1.022, 1.065)	<0.0001	1.057 (1.032, 1.082)	<0.0001
Gaussian Mixture Model	BMI	1.006 (0.957, 1.058)	0.8043	1.009 (0.960, 1.061)	0.7201
	Waist	1.001 (0.982, 1.019)	0.9502	1.002 (0.983, 1.022)	0.8059

* Cluster 2 was categorized as the healthy cluster.

** Adjusted for age and sex.

Table 12 Crude and Adjusted OR (95% CI) based on multinomial logistic regression for 3 derived clusters

Clustering Method	Variable	Crude				Adjusted**			
		Cluster 1 vs. Cluster 3*		Cluster 2 vs. Cluster 3*		Cluster 1 vs. Cluster 3*		Cluster 2 vs. Cluster 3*	
		OR (95% CI)	*p*-value	OR (95% CI)	*p*-value	OR (95% CI)	*p*-value	OR (95% CI)	*p*-value
K-Means	BMI	1.029 (0.943, 1.122)	0.5239	1.304 (1.191, 1.427)	<0.0001	1.035 (0.945, 1.134)	0.4597	1.379 (1.234, 1.542)	<0.0001
	Waist	1.012 (0.984, 1.042)	0.3923	1.096 (1.062, 1.131)	<0.0001	1.015 (0.985, 1.045)	0.3317	1.119 (1.077, 1.163)	<0.0001
Hierarchical Ward's	BMI	1.094 (1.008, 1.189)	0.0324	1.218 (1.133, 1.311)	<0.0001	1.115 (1.020, 1.218)	0.0165	1.259 (1.157, 1.369)	<0.0001
	Waist	1.007 (0.979, 1.036)	0.6130	1.054 (1.030, 1.080)	<0.0001	1.010 (0.981, 1.041)	0.5034	1.065 (1.037, 1.093)	<0.0001
Fuzzy *C*-Means	BMI	1.212 (1.118, 1.314)	<0.0001	1.067 (0.981, 1.161)	0.1319	1.278 (1.164, 1.402)	<0.0001	1.082 (0.990, 1.182)	0.0838
	Waist	1.064 (1.035, 1.093)	<0.0001	1.020 (0.992, 1.048)	0.1574	1.082 (1.049, 1.116)	<0.0001	1.022 (0.994, 1.051)	0.1237

* Cluster 3 was categorized as the healthy cluster.

** Adjusted for age and sex.

CHAPTER V

DISCUSSION

The purpose of this study was to evaluate and compare the performance of 5 clustering algorithms used on 6 inflammatory and obesity biomarkers. *K*-means, Hierarchical Ward's Method, fuzzy *c*-means, and Gaussian Mixture Model clustering algorithms were compared using 4 different validation indices and by their results in the multinomial logistic regression analysis. Although the Principal Component Analysis was not evaluated as the rest of the clustering algorithms, it was also analyzed to determine whether it corroborated the results from the other type of clustering algorithms.

The *k*-means clustering algorithm, followed by Ward's hierarchical agglomerative method, had the best overall performance based on the cluster validation indices used in this study. Specifically, the *k*-means algorithm yielded the highest Calinski-Harabasz and lowest within-cluster sum of squares values, while the hierarchical method had the highest Dunn Index value. However, the statistical analysis showed the *k*-means clustering algorithm with 3 derived clusters had the highest performance compared to the other clustering methods. Out of the three clusters, one of the clusters has the highest levels of CRP and Leptin, another cluster had the highest levels of IL-8, TNFα, and MCP1, and the remaining cluster had the lowest levels for all 6 biomarkers. Logistic regression analysis showed that a higher BMI was associated with the cluster that had the

highest CRP and Leptin levels. Additionally, a higher waist circumference measurement was also associated with the cluster that had the highest CRP and Leptin levels.

Similar findings regarding the positive correlation between CRP and Leptin levels and their association with obesity has also been established in recent studies. Pardina et al. (2010) investigated the levels of IGF-1, CRP, NO, Leptin, and Adiponectin and discovered that CRP and Leptin levels were significantly higher in the 34 morbidly obese patients compared to the 22 normal-weight patients. The obese patients underwent a gastric bypass surgery and 6 months afterwards, CRP levels dropped 57%. Leptin levels did not reduce significantly after 6 months. However, 12 months after the procedure, CRP and Leptin levels in the morbidly obese patients dropped to the same level as the normal-weight patients. Additionally, Valle et al. (2005) studied the difference between CRP, Leptin, and Adiponectin levels in 51 obese children compared to 51 non-obese children, ranging from the age of 6 to 9 years. The authors found that the obese children had significantly higher levels of CRP and Leptin, along with the development of low-grade inflammation. Furthermore, they concluded a positive correlation between log CRP and BMI, along with Leptin and BMI, comparable to our results.

Moreover, researchers have investigated the relation between obesity and inflammation and whether there is any significant connection. Glowinska and Urban (2003) studied the levels of inflammatory and anti-inflammatory cytokines in 64 adolescents, ranging from the age of 12 and 17, with atherosclerosis risk factors, such as obesity, diabetes, and hypertension. The authors found that higher levels of inflammatory cytokines, including IL-6 and TNFa, are positively correlated with BMI in obese adolescents and slim adolescents with hypertension. This demonstrates that there could be a presence of low-grade inflammation before reaching an obesity weight level.

Similar conclusions were found in the Utsal et al. (2012) were 13 biomarkers were studied in 76 boys, ages 10 to 11 years old, half of which had a normal BMI and the rest had an overweight BMI. Researchers found significantly higher levels of 6 of the 13 biomarkers, including IL-6, IL-8, MCp1, and CRP, in overweight boys compared to the boys with a normal BMI. These studies demonstrate that the presence of inflammation is still possible in young adults who are overweight although not obese. The evidence presented in these studies support the idea that although inflammation can certainly be present in individuals who are obese, those who are overweight may also exhibit mild inflammation. This distinction is significant because mild inflammation in an adolescent may be a potential risk factor for developing obesity in later years.

It should be noted that clustering algorithms will always generate clusters, even if there is no meaningful underlying structure in the data set. Therefore, it is crucial to evaluate the resulting clusters to determine their actual statistical significance. While Ward's hierarchical agglomerative method and fuzzy *c*-means algorithms yielded clinically relevant clusters similar to *k*-means, they received poor evaluations from the cluster validation indices. On the contrary, the *k*-means algorithm not only yielded clinically relevant clusters, but also held the highest rank based on the validation indices. Prior research has also demonstrated *k*-means to be one of the most efficient clustering algorithms. Velmurugan (2014) compared the performance of *k*-Means and fuzzy *c*-means clustering algorithms using a 12 telecommunication data sets and concluded *k*-Means was more favorable due to its faster computation time and evenly distributed clusters. However, Dubey et al, (2018) also compared the performance of *k*-means and FCM using only 699 records and found FCM had a 97% accuracy rate as opposed to 92% from the *k*-means method. This illustrates the impact the size of a data set can have on the outcomes of the clustering algorithms. Other common types of clustering algorithms include Hierarchical methods. Zhao & Karypis (2002)

analyzed 12 different datasets, ranging from 878 to 4069 data points, using partitional and hierarchical methods. They concluded partitioning methods were more accurate than hierarchical methods, especially when working with bigger data sets. Since Hierarchical clustering algorithms are computationally intensive, when dealing with larger data sets it is more prone to errors resulting in a higher inaccuracy (Gülagiz & Sahin, 2017).

Furthermore, Ahmad & Dang (2015) evaluated partitional, hierarchical, and density-based clustering algorithms using different small and large data sets. The data sets ranged from 8 to 1554 attributes and 100 to 2924 observations. The researchers concluded that the *k*-means clustering algorithm was the most efficient and effective due to its low computational complexity and ability to handle both small and relatively large datasets. Some of the limitations for the hierarchical and density-based algorithms discussed in their study were time complexity, sensitivity to outliers, and difficulty handling large data sets. In Kameshwaran & Malarvizhi (2014), the researchers noted that one of the advantages of hierarchical and density-based models is that a-priori specification, such as the number of clusters computed, is not necessary. As for the disadvantages, they noted one major issue with the hierarchical clustering algorithm is that it cannot revise or undo previously computed clusters, which may lead to inaccurate clusters if incorrect assumptions are made in the early computational stages.

Additionally, in our study we conducted Principal Component Analysis, where 2 principal components were extracted. However, the two components did not account for more than 80% of the variance. The first principal component included CRP and Leptin, while the second principal component included HGF, IL-8, MCP1, and TNFα. Although the explained variance ratio was not

met, the analysis supported the variables found in each component were similar to the variables clustered together in 2 of the groups that were determined by the *k*-means algorithm.

Overall, the *k*-means clustering algorithm was found to produce the most accurate clusters for obesity-inflammatory biomarkers considering several factors. Cluster validation indices indicated that it performed favorably compared to the other algorithms based on half of the indices applied. Furthermore, the similarity of the resulting clusters to the principal components computed by PCA provided further evidence. Additionally, when the clusters were interpreted in a clinical context, they were found to be relevant and significant.

While our study has yielded valuable insights, it is not without limitations. The sample size was relatively small, consisting of only 183 participants. After pre-processing the data for analysis, all subjects with missing data points were removed, leading to a reduced sample size of 163 participants. Although this type of limitation could reduce the statistical power of the analysis, the clustering algorithms were still able to identify clusters of inflammatory biomarkers that are supported by other research studies. Another limitation was that the participants used for our sample were from a local adolescent population. Therefore, our findings cannot be extended to the general population. Additionally, the study utilized a cross-sectional design, which poses challenges in establishing causality or exploring the changes over time between variables. Although the outcomes indicate associations among variables in a particular instance, it is insufficient for determining causation. Overall, these limitations should be kept in mind when interpreting the study's findings and caution should be exercised when extending the conclusions to other populations.

Clustering algorithms have become increasingly popular with the rise of data due to their wide range of applications across various fields. While having various clustering techniques available, it can become challenging determining which method is best suited for a particular study. Abbas (2008) compares the performance of *k*-means, Hierarchical, Self-Organizing Maps, and Expectation Maximization clustering algorithms and concluded *k*-means and EM algorithms performed better than the Hierarchical algorithm overall. Their performance significantly increases when using big data, however, while using a small data set, the Hierarchical clustering algorithm outperformed the other methods. Data sets can vary in size or types and not all algorithms will be equally effective for all. Selecting an inappropriate algorithm can lead to inaccurate results or misinterpretation of the data. It is essential to carefully evaluate the characteristics of the data and the limitations of different clustering algorithms.

While clustering algorithms can be effective in recognizing patterns in data, it ultimately is up to the researcher to correctly interpret the derived clusters accurately since the algorithms will produce clusters regardless of significance. This may become an issue since the interpretation of the results will depend on the researcher's level of knowledge in the field. Misinterpretation of data can lead to incorrect conclusions and potentially harmful decisions. Therefore, it is important for researchers to continue improving clustering interpretability.

REFERENCES

Abbas, O. A. (2008). Comparisons between data clustering algorithms. *International Arab Journal of Information Technology (IAJIT)*, *5*(3).

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433-459.

Abraham, B., & Ledolter, J. (2006), Introduction to Regression Modeling, Thomson Brooks/Cole, Belmont, CA.

Ahmad, H. P., & Dang, S. (2015). Performance Evaluation of Clustering Algorithm Using different dataset. *International Journal of Advance Research in Computer Science and Management Studies*, *8*.

Allam, A., & Gumpeny, R. S. (2012). Analyzing microarray data of Alzheimer's using cluster analysis to identify the biomarker genes. *International journal of Alzheimer's disease*, *2012*.

Amin, M. N., Hussain, M. S., Sarwar, M. S., Moghal, M. M. R., Das, A., Hossain, M. Z., ... & Islam, M. S. (2019). How the association between obesity and inflammation may lead to insulin resistance and cancer. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *13*(2), 1213-1224.

Bair, E. (2013). Semi-supervised clustering methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(5), 349-361.

Befort, C. A., Nazir, N., & Perri, M. G. (2012). Prevalence of obesity among adults from rural and urban areas of the United States: findings from NHANES (2005-2008). *The Journal of Rural Health*, *28*(4), 392-397.

Bezdek, J. C., Hathaway, R. J., Sabin, M. J., & Tucker, W. T. (1987). Convergence theory for fuzzy c-means: counterexamples and repairs. *IEEE Transactions on Systems, Man, and Cybernetics*, *17*(5), 873-877.

Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). clValid: An R package for cluster validation. *Journal of Statistical Software*, *25*, 1-22.

Brownell, K. D. (2010). The humbling experience of treating obesity: Should we persist or desist?. *Behaviour Research and Therapy*, *48*(8), 717-719.

Chi-Hsien, K., & Nagasawa, S. (2019). Applying machine learning to market analysis: Knowing your luxury consumer. *Journal of Management Analytics*, *6*(4), 404-419.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1-22.

Dornbush, S., & Aeddula, N. R. (2021). Physiology, Leptin. In *StatPearls [Internet]*. StatPearls Publishing.

Dubey, A. K., Gupta, U., & Jain, S. (2018). Comparative study of K-means and fuzzy C-means algorithms on the breast cancer data. *International Journal on Advanced Science, Engineering and Information Technology*, *8*(1), 18-29.

El-Habil, A. M. (2012). An application on multinomial logistic regression model. *Pakistan journal of statistics and operation research*, 271-291.

Ellulu, M. S., Patimah, I., Khaza'ai, H., Rahmat, A., & Abed, Y. (2017). Obesity and inflammation: the linking mechanism and the complications. *Archives of medical science*, *13*(4), 851-863.

Gan, H., Sang, N., Huang, R., Tong, X., & Dan, Z. (2013). Using clustering analysis to improve semi supervised classification. *Neurocomputing*, *101*, 290-298.

Gentleman, R., & Carey, V. J. (2008). Unsupervised machine learning. In *Bioconductor case studies* (pp. 121-157). Springer, New York, NY.

Głowińska, B., & Urban, M. (2003). Selected cytokines (Il-6, Il-8, Il-10, MCP-1, TNF-alpha) in children and adolescents with atherosclerosis risk factors: obesity, hypertension, diabetes. *Wiadomosci Lekarskie (Warsaw, Poland: 1960)*, *56*(3-4), 109-116.

Gülagiz, F. K., & Sahin, S. (2017). Comparison of hierarchical and non-hierarchical clustering algorithms. *International Journal of Computer Engineering and Information Technology*, *9*(1), 6.

Hammouda, K., & Karray, F. (2000). A comparative study of data clustering techniques. *University of Waterloo, Ontario, Canada*, *1*.

Hennig, C., & Imports, M. A. S. S. (2015). Package 'fpc'. *Flexible Procedures for Clustering.*

Hiran, K. K., Jain, R. K., Lakhwani, K., & Doshi, R. (2021). *Machine Learning: Master Supervised and Unsupervised Learning Algorithms with Real Examples (English Edition)*. BPB Publications.

Hossain, P., Kawar, B., & El Nahas, M. (2007). Obesity and diabetes in the developing world—a growing challenge. *New England journal of medicine*, *356*(3), 213-215.

Hotamisligil, G. S. (2006). Inflammation and metabolic disorders. *Nature*, *444*(7121), 860-867.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264-323.

Kaufman, L. & Rousseeuw, P.J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis, Wiley.

Kameshwaran, K., & Malarvizhi, K. (2014). Survey on clustering techniques in data mining. *International Journal of Computer Science and Information Technologies*, *5*(2), 2272-2276.

Kanda, H., Tateya, S., Tamori, Y., Kotani, K., Hiasa, K. I., Kitazawa, R., ... & Kasuga, M. (2006). MCP-1 contributes to macrophage infiltration into adipose tissue, insulin resistance, and hepatic steatosis in obesity. *The Journal of clinical investigation*, *116*(6), 1494-1505.

Kassambara, A., & Mundt, F. (2017). Package 'factoextra'. *Extract and visualize the results of multivariate data analyses*, *76*(2).

Kelly, I. R., Grossman, M., & Chou, S. Y. (2005). The super size of America: an economic estimation of body mass index and obesity in adults.

Kim, C. S., Park, H. S., Kawada, T., Kim, J. H., Lim, D., Hubbard, N. E., ... & Yu, R. (2006). Circulating levels of MCP-1 and IL-8 are elevated in human obese subjects and associated with obesity-related parameters. *International journal of obesity*, *30*(9), 1347-1355.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory*, *28*(2), 129-137.

Lumeng, C. N., & Saltiel, A. R. (2011). Inflammatory links between obesity and metabolic disease. *The Journal of clinical investigation*, *121*(6), 2111-2117.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., Studer, M., ... & Gonzalez, J. (2013). Package 'cluster'. *Dosegljivo na.*

MacQueen, J. (1967). Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability* (pp. 281-297).

Masters, R. K., Powers, D. A., & Link, B. G. (2013). Obesity and US mortality risk over the adult life course. *American journal of epidemiology*, *177*(5), 431-442.

McAlpine, E. D., Michelow, P., & Celik, T. (2022). The utility of unsupervised machine learning in anatomic pathology. *American Journal of Clinical Pathology*, *157*(1), 5-14.

McArdle, M. A., Finucane, O. M., Connaughton, R. M., McMorrow, A. M., & Roche, H. M. (2013). Mechanisms of obesity-induced inflammation and insulin resistance: insights into the emerging role of nutritional strategies. *Frontiers in endocrinology*, *4*, 52.

Moore, W. C., Meyers, D. A., Wenzel, S. E., Teague, W. G., Li, H., Li, X., ... & Bleecker, E. R. (2010). Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. *American journal of respiratory and critical care medicine*, *181*(4), 315-323.

Nwogbaga, N. E. (2020). A review of big data clustering methods and research issues. *Int J Sci Res (IJSR)*, *9*(5), 253-264.

Ogden, C. L. (2007). *Obesity among adults in the United States-no change since 2003-2004* (No. 1). US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

Palacio-Niño, J. O., & Berzal, F. (2019). Evaluation metrics for unsupervised learning algorithms. *arXiv preprint arXiv:1905.05667*.

Pardina, E., Ferrer, R., Baena-Fustegueras, J. A., Lecube, A., Fort, J. M., Vargas, V., ... & Peinado-Onsurbe, J. (2010). The relationships between IGF-1 and CRP, NO, leptin, and adiponectin during weight loss in the morbidly obese. *Obesity surgery*, *20*, 623-632.

Park, H. S., Park, J. Y., & Yu, R. (2005). Relationship of obesity and visceral adiposity with serum concentrations of CRP, TNF-α and IL-6. *Diabetes research and clinical practice*, *69*(1), 29-35.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, *2*(11), 559-572.

Poirier, P., Giles, T. D., Bray, G. A., Hong, Y., Stern, J. S., Pi-Sunyer, F. X., & Eckel, R. H. (2006). Obesity and cardiovascular disease: pathophysiology, evaluation, and effect of weight loss: an update of the 1997 American Heart Association Scientific Statement on Obesity and Heart Disease from the Obesity Committee of the Council on Nutrition, Physical Activity, and Metabolism. *Circulation*, *113*(6), 898-918.

Santos, A. C., Lopes, C., Guimaraes, J. T., & Barros, H. (2005). Central obesity as a major determinant of increased high-sensitivity C-reactive protein in metabolic syndrome. *International journal of obesity*, *29*(12), 1452-1456.

Scherzer, R., Shah, S. J., Secemsky, E., Butler, J., Grunfeld, C., Shlipak, M. G., & Hsue, P. Y. (2018). Association of biomarker clusters with cardiac phenotypes and mortality in patients with HIV infection. *Circulation: Heart Failure*, *11*(4), e004312.

Serdula, M. K., Ivery, D., Coates, R. J., Freedman, D. S., Williamson, D. F., & Byers, T. (1993). Do obese children become obese adults? A review of the literature. *Preventive medicine*, *22*(2), 167-177.

Shi, P., & Goodson, J. M. (2019). A data mining approach identified salivary biomarkers that discriminate between two obesity measures. *Journal of Obesity*, *2019*.

Shlens, J. (2014). A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE access*, *8*, 8071680727.

Syarif, I., Prugel-Bennett, A., & Wills, G. (2012). Unsupervised clustering approach for network anomaly detection. In *International conference on networked digital technologies* (pp. 135-145). Springer, Berlin, Heidelberg.

Tönnies, T., Brinks, R., Isom, S., Dabelea, D., Divers, J., Mayer-Davis, E. J., ... & Imperatore, G. (2023). Projections of type 1 and type 2 diabetes burden in the US population aged< 20 years through 2060: the SEARCH for Diabetes in Youth study. *Diabetes Care*, *46*(2), 313-320.

Utsal, L., Tillmann, V., Zilmer, M., Mäestu, J., Purge, P., Jürimäe, J., ... & Jürimäe, T. (2012). Elevated serum IL-6, IL-8, MCP-1, CRP, and IFN-γ levels in 10-to 11-year-old boys with increased BMI. *Hormone research in paediatrics*, *78*(1), 31-39.

Valle, M., Martos, R., Gascon, F., Canete, R., Zafra, M. A., & Morales, R. (2005). Low-grade systemic inflammation, hypoadiponectinemia and a high concentration of leptin are present in very young obese children, and correlate with metabolic syndrome. *Diabetes & metabolism*, *31*(1), 55-62.

Velmurugan, T. (2014). Performance based analysis between k-Means and Fuzzy C-Means clustering algorithms for connection oriented telecommunication data. *Applied Soft Computing*, *19*, 134-146.

Villarín, M. C. (2019). Methodology based on fine spatial scale and preliminary clustering to improve multivariate linear regression analysis of domestic water consumption. *Applied Geography*, *103*, 22-39.

Visser, M., Bouter, L. M., McQuillan, G. M., Wener, M. H., & Harris, T. B. (2001). Low-grade systemic inflammation in overweight children. *Pediatrics*, *107*(1), e13-e13.

Wang, Y., Beydoun, M. A., Liang, L., Caballero, B., & Kumanyika, S. K. (2008). Will all Americans become overweight or obese? Estimating the progression and cost of the US obesity epidemic. *Obesity*, *16*(10), 2323-2330.

Wang, Z., & Nakayama, T. (2010). Inflammation, a link between obesity and cardiovascular disease. *Mediators of inflammation*, *2010*.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks*, *16*(3), 645-678.

Yim, O., & Ramdeen, K. T. (2015). Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. *The quantitative methods for psychology*, *11*(1), 8-21.

Zerhari, B., Lahcen, A. A., & Mouline, S. (2015). Big data clustering: Algorithms and challenges. In *Proc. of Int. Conf. on Big Data, Cloud and Applications (BDCA'15)*.

Zhao, Y., & Karypis, G. (2002, November). Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the eleventh international conference on Information and knowledge management* (pp. 515-524).

Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked*, *17*, 100179.

BIOGRAPHICAL SKETCH

Tania Mayleth Vargas graduated from The University of Texas at Austin with a Bachelor of Science and Arts in Biology and minor in Education and Curriculum in December 2019. She continued her education at The University of Texas Rio Grande Valley and earned her Master of Science in Applied Statistics and Data Science in May 2023. She can be contacted via email at taniamaylethvargas@gmail.com.