TECHNICAL UNIVERSITY OF MUNICH

TUM Data Innovation Lab

Defining Corporate Health Classes

Authors	Anastasia Kireeva Qinyan Li Roxana Mirshahvalad Felix Sievers Nora Walkembach
Mentor(s)	Dr. Dominik Jüstel (Helmholtz Zentrum) M.Sc. Jan Kukačka (Helmholtz Zentrum) Dr. Sebastian Dünnebeil (Wellabe)
Co-Mentor	M.Sc. Stefan Bamberger
Project Lead	Dr. Ricardo Acevedo Cabra (Department of Mathematics)
Supervisor	Prof. Dr. Massimo Fornasier (Department of Mathematics)

Jul 2020

Abstract

The importance of taking care of your well-being is becoming increasingly prominent in society. As a result, we have seen the health and fitness industry growing continuously in recent years. Corporations have also started to realize the importance of healthy employees to an efficient and successful running of their company. Wellabe is a start-up which is helping their customer companies achieve better overall health of their workforce through on-site check-ups, modern video consultation and digital prevention programs. The individualized digital programs are based on the data collected during the health check-ups. Defining health classes is necessary to determine for whom which program is suited best. The goal of this project is to do this based on Wellabe's dataset of the check-ups. Due to the evident privacy concerns that come with medical data the original dataset could not be shared with us. Instead, we got the unique opportunity to work with a synthetically generated version of the dataset. This not only allowed us to explore what health classes we could define but also the potential and limitations of using synthetically generated data to find patterns in the original data.

We started our project by conducting a detailed analysis of the dataset to get a comprehensive understanding of the underlying structures in the data and the medical significance of outliers. The data is pre-processed based on the gained domain knowledge. A score which was given by a doctor as a judgement of one's health status and the difference between one's biological and chronological age are used as indicators of general health. The biological age is estimated based on the chronological age in the dataset. The doctor's score and the biological age are predicted using Multiple Linear Regression (MLR), Generalized Additive Models (GAM) and Multivariate Adaptive Regression Splines (MARS). The best prediction was achieved using GAMs for biological age and MARS for the doctor's score. Significantly better predictions were achieved using the original data rather than the synthetically generated one showing that there are clear limitations in using synthetic data.

To define health classes, the dataset was clustered using K-means, Density Clustering and Hierarchical Clustering. Features are selected based on the findings in our medical analysis and how dominant they were in the previous predictions. Out of the three explored clustering methods only Hierarchical Clustering was able to define valuable health classes. These resulting classes are not only able to separate the healthy from the unhealthy people but also can identify people with similar health issues such as fatty liver disease and obesity. These clusters were additionally able to show that the common patterns replicated in the synthetic data were sufficient enough for the successful clustering of medically relevant groups. The hierarchical nature of the clustering also provides opportunities to divide and explore the classes in further detail.

Contents

1 Introduction

1.1 Problem Definition and Goals of the Project

Wellabe is a start-up which enables employees of their customer's companies to better understand their health and actively improve it with individualized recommendations made possible through health check-ups. These check-ups are conducted at their clients' companies and consist of taking heart measurements, analysing a few blood drops, measuring body composition values through a smart scale, taking controlled measurements of a few of breaths and testing the extent of one's mobility. All measured data is examined by health experts who then give individualized scores based on the state of one's health. All this information acquired makes up Wellabe's unique health dataset which has much potential for exciting data explorations and analysis.

Wellabe offers tailored prevention plans and programs through their mobile app. To determine for whom which program or plan is best suited, health classes need to be defined. This is where we came in. Our main goal for this project is to find a way to define relevant health classes, which have the potential to improve the individualized recommendations. These are the four main steps we came up with to achieve this goal:

1. **Data Understanding and Handling**
   (a) Detailed analysis of the dataset to give a comprehensive understanding of the underlying structures in the data and the medical significance of outliers
   (b) Preprocessing based on domain knowledge
2. **Finding Indicators of General Health**
   (a) Identify potential values that can indicate one's overall health
   (b) Predict the indicators in the given dataset
   (c) Select the features which are most dominant in the prediction for the final clustering
3. **Defining health classes**
   (a) Use unsupervised clustering to find clusters in the dataset
   (b) Conduct a detailed medical analysis of the clusters

Due to the exploratory nature of this project, we had the freedom to define two relevant side objectives to investigate in more detail:

1. **Age Prediction**
   Once we decided to use the difference between biological and chronological age as an indicator of general health we were told by our partner company Wellabe that they would like us to explore biological age prediction in more detail as it could be of great value to them. This is due to the fact that they display a calculated biological age to their users in the mobile app. Hence, we decided to make the prediction of the biological age a separate side objective of this project.

2. **Limitations and potential of synthetically generated data**
   As clinical data is very personal, sharing it freely is not an option without guaranteed anonymity. The use of generated synthetic data to deal with problems as such is of very recent research. Hence not much is known about how it affects data insights compared to the ones made from the original data. With this unique opportunity, it only made sense to spend some time looking at how well different algorithms performed on the real data versus the synthetic data.

1.2 State of the Art Approaches and Algorithms

1.2.1 Synthetically Generated Medical Data

Working with medical dataset comes with many challenges. Two major ones being security and outlier handling. One's medical data is something very personal and thus should at no cost be given to third parties. Synthetically generating data which contains none of the individual's information but yet maintains all the underlying structures would solve this security issue. However, achieving this is far from straightforward but there is extensive ongoing research into improving synthetic data generation. Loong et al. have concluded that synthetic data might be best suited for preliminary data analysis purposes for now [9]. The challenge with outliers in medical data is that they can have many reasons including measurement errors and illnesses. This can be a problem as one might want to keep the data of people with naturally abnormal values but remove all the errors caused by measurement. Differentiating between those can become a critical issue. Using domain knowledge to define limits is a tedious but safe way to ensure only extremely improbable values are removed.

1.2.2 Biological Age

Ageing has long been known to be correlated to general health. In 1996, Chodzko-Zajko [3] published a paper on the inevitability of the decline of health markers with increasing age. The rate of decline of those health markers differs from person to person. One's current state is often referred to as one's biological age by scientists [13]. One can estimate the biological age by predicting the chronological age from a large enough dataset of mostly healthy subjects and relevant biomarkers. This has mostly been done using Multiple Linear Regression Models due to the often linear decline of health markers with age [5].

1.2.3 Clustering Medical Data

Methods such as clustering are frequently used to reveal hidden structures and groups in large datasets. For clinical and health dataset, clustering helps to group the data and characterize differences between objects. Hirano et al. [7] tried hierarchical clustering on clinical databases and the best clusters were obtained using Ward's method where the clinically reasonable attributes were selected. Some other clustering methods such as kmeans were also shown to work on some medical dataset [8].

2 Data Understanding and Handling

The given dataset contains data points from fifty thousands different subjects where 57.5% are male and 42.5% are female and it has 70 features. The features in the dataset can be put into four main categories of numerical bio-markers with some additional features like some indicators of user's behavior such as 'ate_recently' and the review score, which is a integer score given by a doctor defining the health status of a person. The four main categories are Cardiovascular System, Metabolism, Respiratory System and Body Composition. The Cardiovascular system is a system of organs that permits blood to circulate and transport nutrients through the body. Metabolism describes all the chemical reactions inside the body required to make the organs function properly. The Respiratory System consists of a group of organs responsible for absorbing oxygen from the inhaled air and expelling carbon dioxide through the exhaled air. The Body Composition features contain general body information such as height and weight and others were used to describe the percentages of fat, bone, water and muscle in human bodies.

Features of these main categories are shown in Table 1.

Cardiovascular System	Metabolism	Respiratory System	Body Composition
Systolic Blood Pressure	Blood Sugar	Oxygen Saturation	Body Weight
Diastolic Blood Pressure	Triglycerides	Forced Vital Capacity	Visceral Fat Level
Pulse Pressure	Cholesterol	Forced Expiratory Volume in 1 Second	Daily Caloric Needs
Mean Arterial Pressure	High-density lipoproteins	Peak Expiratory Flow	Muscle Mass
Ankle-Brachial Index	Alanine transaminase	Forced Expiratory Flow at 25% of the Lung Volume	Body Fat Mass
Resting Heart Rate	Aspartate transaminase		Body Fat Percentage
	Gamma-glutamyl transferase		Body Water Percentage
	Creatine		Bone Mineral Mass
	Fat liver index		Body Height

Table 1: List of features of four main parts of data-set

2.1 Statistical Data Exploration

We started by familiarizing ourselves with the medical meaning behind the values of each feature, finding their normal ranges and finding diseases which are related to values outside of the normal. The definition of features and their normal ranges are shown in table 4 and 5.

Next, we started by looking at the descriptive statistics of the dataset such as mean, median and standard deviation of each feature. This helped us to identify inconsistencies in the dataset. For example, all of the features in the dataset should have positive values. By looking at the minimum value of each feature, we can identify if there exist any negative value.

Afterwards, we looked at the distribution of features to find out which features were normally distributed and which were not. For this reason, we used histograms of distribution of features and normal probability plots. Normal probability plot indicates normal distribution if the distribution of data follows closely the diagonal line. Then, we investigated the relationship between features to find out if there exist a linear relationship between features. To do this, we looked at the scatter plot of continuous features to gain a first impression. Since scatter plot does not indicate the strength of relationship amongst features, we also used sample Pearson correlation coefficients (r_{xy}) to find out which features are highly correlated. Sample Pearson correlation coefficient is calculated using:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} \tag{1}$$

where s_{xy} is sample covariance and s_x and s_y are sample standard deviations. Sample Pearson correlation coefficient varies between $[-1, +1]$. -1 indicates perfect negative linear correlation, +1 indicates perfect positive linear correlation and 0 indicates no correlation between two features. The last two important steps were handling missing values and detecting outliers. Outliers can affect our models and can be a valuable source of information, providing us insights about specific behaviours. For example in Metabolism features we have very high values which at first glance they seemed to be outliers, however after more investigation we realized that they were indication of diseases and we should not consider them as outliers. These two steps will be explained in more details in the section 2.2.

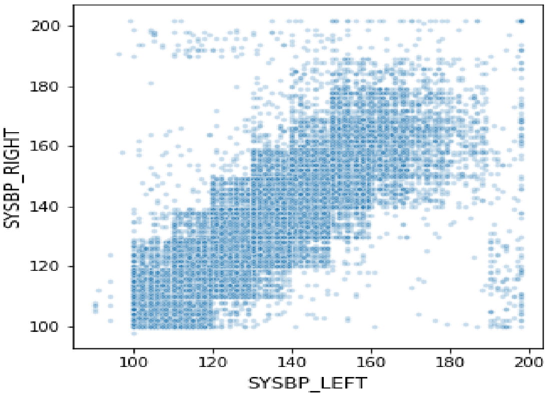
Due to the previously mentioned privacy concerns with clinical data, it was not possible to share the original health data with us in this project. Hence, the dataset given to us was synthetically generated based on the original data guaranteeing anonymity for each individual in the original dataset. This artificially generated synthetic dataset is replicating specific structures of the real dataset such as the distribution of all features and the relationship between them. Although synthetic data has many advantages such the guaranteed data privacy, it also comes with some limitations.

Models which generate synthetic data mostly identify common trends in the original data, but subtle relationships might be missed. Additionally, this creates susceptibility to statistical noise and it may lead to undesired synthetic patterns in the generated dataset. During data exploration we noticed some of these synthetic patterns. In figure 1 which is a scatter plot of systolic blood pressure left (SYSBP_Left) versus systolic blood pressure right (SYSBP_right), Rectangular patterns can clearly be identified. Some of these generated patterns may introduce some illogical and improbable values to the dataset. Another challenge presented by the use of synthetic data is the need for verification of one's findings, meaning that performing identical analysis on the original data to test and compare results is often necessary. This ensures that the system has been properly trained and is not generating outputs based on any incorrect artificial patterns generated in the synthetic data.

2.2 Data Cleaning

In order to improve the quality of the given dataset, unexpected, incorrect, and inconsistent data points were detected and replaced by a NaN value. Afterwards, we provided

Figure 1: Rectangular patterns found in scatter plot of SYSBP_Left vs. SYSBP_Right



the possibility to impute missing values using feature specific methods, so that we could have more information for further analysis .

Detecting errors in medical data can be challenging as extreme values could either be an indication of a disease or a measurement error. Since these outliers may greatly affect the final result, it is necessary to remove them with caution to avoid the elimination of values from sick people. Thus we carefully defined thresholds by also considering the relationships between features. The detected outliers were replaced by NaNs. Additionally, we tried to find reasonable estimations of the NaN values based on the medians, formulas from literature and other methods such as linear regression.

As explained in section 2.1, the given dataset has features from four main categories. Features of each category have their own individual set of properties and traits, we separately processed each feature of each section. For instance, in Metabolism section, Total Cholesterol is calculated by adding high-density lipoprotein (HDL), low-density lipoprotein (LDL) and 20% of Triglycerides (TG). So, it is not possible for someone to have less than or equal amount of cholesterol as HDL plus 20% of TG. Therefore, values which did not comply with this were replaced by NaN.

Additionally, to estimate the missing values of for instance the body water percentage (BWP) from Body Composition section, which is calculated based on other features such as age, height and weight of a person, the Watson formula from the literature [15] was used. To estimate BWP, the total body weight (TBW) which is calculated differently for men and women was calculated. Then, TBW will be divided by the weight of the person to get the percentage. The formulas of calculating BWP for male are shown below:

$$TBW(male) = 2.447 - (0.09156 \times age) + (0.1074 \times height) + (0.3362 \times weight) \quad (2)$$

$$BWP = \frac{TBW}{weight} \times 100\% \quad (3)$$

Detailed preprocessing of each feature in all sections can be seen in table 6 and 7 in appendix.

3 Finding Indicators of General Health

We discovered two methods for defining a potential indicator of general health which we decided to investigate in great detail. Firstly, we predicted the age based on the other biomarkers, which gave us either an over- or underestimation of the real age. This deviation can be interpreted as an indication of general health. Secondly, we predicted the review scores given by health experts which allowed us to identify the most contributing features to this prediction that could then be selected for clustering. Both predictions are performed using three different methods which are discussed in the following section.

3.1 Regression Methods

3.1.1 Multiple Linear Regression (MLR)

In multiple linear regression models we assume that the dependent variable y can be described by a set of linear dependencies from independent variables x_i. The model function has the form of

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{4}$$

where β_0 is the offset and the other β_i are the slopes according to each variable x_i. The fitting is done via ordinary least squares,

$$\min_{\boldsymbol{\beta}} \|X\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 \tag{5}$$

where X is a matrix consisting of all the features from the data that the model is trained on $(\boldsymbol{x}_1 ... \boldsymbol{x}_k)$, $\boldsymbol{\beta}$ is a vector containing all the slopes β_i and $\boldsymbol{y}$ is the solution vector.

3.1.2 Multivariate Adaptive Regression Splines (MARS)

Multivariate Adaptive Regression Splines builds a piecewise linear model using functions of the form $(x-t)_+ := \max(0, x-t)$ or $(t-x)_+$, where x is a feature value, and t is a threshold. These functions are called hinge functions with a knot of the value t. Two functions with the same feature X_j and threshold t, but different sign of difference $(X_j - t)_+, (t - X_j)_+$ are called a reflected pair.

The model has the form $\tilde{f}(X) = \beta_0 + \sum_{k=1}^{K} \beta_k h_k(X)$, where $h_k(X)$ is a hinge function or a product of two or more hinge functions.

The model building procedure consists of two stages: forward model-building procedure and backward deletion procedure. During the forward pass new terms producing the largest decrease in a training residual square error are added iteratively. The considered terms are of the form

$$\beta_{M+1} h_l(X)(X_j - t)_+ + \beta_{M+2} h_l(X)(t - X_j)_+ \tag{6}$$

where $h_l(X)$ is one of already selected functions and $(X_j - t)_+, (t - X_j)_+$ is a candidate reflected pair.

To prevent overfitting after the forward pass, the backward deletion procedure is applied. At each step the term whose removal causes the smallest increase in error is deleted. To compare performance of the models with different subsets of terms, generalized cross validation is used:

$$GCV(\lambda) = \frac{\sum_{i=1}^{N}(y_i - f_\lambda(x_i))^2}{(1 - M(\lambda)/N)^2}, \tag{7}$$

where λ is a number of terms, and $M(\lambda)$ is an effective number of parameters in the model.

We decided to use MARS as it can identify the most important features along with their interactions and the important thresholds. The advantage of using MARS is that it is able to operate locally as the hinge function or the product of hinge functions is zero over part of their range. MARS can also learn non-linear dependencies which methods like Multivariate Linear Regression can not. Unlike Regression Tree which uses $I(x - t > 0)$ and $I(x - t \leq 0)$, MARS additionally accounts for the difference between the value and the threshold which is especially relevant when using medical data.

3.1.3 Generalized Additive Models (GAM)

A GAM is a generalized linear model. It has an additive structure of smooth functions of the form [1]

$$g(\mathbb{E}[y|X])) = \beta_0 + f_1(X_1) + f_2(X_2) + \ldots + f_M(X_N) \tag{8}$$

where g is a linking function, and f_i are feature functions that are built using penalized B splines. They use the data points as controls points and create smooth curves by combining flexible functions from point to point. Furthermore, it is assumed that the dependent variable y given parameter X has a distribution from an exponential family. An exponential family is a parametric set of probability distributions. Examples are the family of normal, binomial or Poisson distributions. The feature functions allow to automatically model non-linear relationships without manually transforming each feature. This is especially relevant for features which are not normally distributed, like the ones given in this dataset. The functions are fitted to the data via minimizing the negative likelihood with an additional integral over the second derivatives of the feature functions. This keeps the functions smooth and prevents them from overfitting. The minimisation problem is

$$\min\left(-L(x_1, \ldots, x_n) + \sum_i \lambda_i \int f_i''(x)^2 dx\right). \tag{9}$$

where L is the likelihood function and λ_i are the penalty parameters and have to be tuned. This is usually done via grid search.

Essentially, the model consists of three different parts:

1. the distribution from $y \sim$ Exponential Family$(\mu \mid X)$

2. the linking function g
3. the functional form with an additive structure

For the purpose of this project, the distribution was chosen to be a gamma distribution, as we have purely positive values and cannot tell whether y given X is normally distributed due to the high dimensionality of the dataset. The linking function was chosen to be the identity function as there is no need to increase the complexity any further.

3.2 Age Prediction

The purpose of predicting the biological age was to have a marker of one's physiological health. Individuals with an estimated age close to their real age are expected to have similar values to their average healthy colleagues in the same age group. The lower the predicted age relative to the real age, the more healthy the individual is compared to the average and vice versa for higher predicted age. For this to be true, the younger the person is the healthier they should be on average, which is true for this dataset. As we expect unhealthy or even sick people to have a predicted age above their real age and very healthy and fit people to have a predicted age below their real age, it makes sense to use the difference between the real and the estimated age as an indicator of one's general health.

3.3 Review Score Prediction

The so-called Review Score is a discrete score from zero to six given by a doctor based on the results of the checkup. A review score of six is given to individuals of optimal health, and a score of zero should be given in the case of a medical emergency where one is encouraged to consult a doctor as soon as possible. However, in the beginning, the score zero was misused as it was the only way to leave a personal note for the patient. Therefore, the rows with a review score equal to zero are not reliable and were removed.

Since the score is given by a professional based on the checkup results it can be interpreted as a health indicator. We also saw potential in being able to automatically predict the review score without a consultation with a doctor.

4 Defining Health Classes

The main goal of the project is to define medically relevant health classes. The idea is to group people with similar health conditions together based on the results of their checkup. Having groups of people with similar health markers allow for the determination of possible risks and hence the ability to give personalized recommendations.

As the relations between the features of the dataset are rather complex and mostly continuous, we did not expect to obtain a clear and distinct division of all instances. It is also worth noting that the defined health classes are expected to vary in size when considering that they should be based on various health conditions which are not all equally common. Ideally, we want to identify small groups of people with a high risk for a certain disease and larger groups containing people that live generally healthy or unhealthy lifestyles.

We implemented several clustering algorithms (reviewed in Section 4.1) and tried to compare the results to be able to select the most interpretable groups for further exploration.

4.1 Clustering Methods

Clustering algorithms divide data in a way that achieves high similarities within clusters and low similarities between clusters, hence finding natural groupings among the instances. Based on some existing research on medical data such as in Section 1.2, we tried several clustering methods on our health dataset which are described below.

4.1.1 K-means Clustering

K-means is a simple, easily interpretable clustering algorithm which can be used to group a dataset into a predefined number of clusters. It was commonly used on health data [12, 16] in existing research work. It initializes the centroids (one per cluster) and aims to minimize the within-cluster sum of squares (i.e. variance) by recursively optimizing the objective and updating the centroids of all clusters until convergence.

4.1.2 Density-based Clustering

Density-based Clustering identifies clusters by finding dense areas of points separated by sparse regions of low density based on a density condition. It does not require the number of clusters as it automatically infers it from the data.

A well-known implementation of this archetype of clustering is DBSCAN. Its general idea is to continue growing a given cluster as long as the density in a certain neighbourhood exceeds a certain threshold.

4.1.3 Hierarchical Clustering

Unlike K-means and Density-based Clustering, Hierarchical Clustering is not just segmenting the data space once but intends to create a hierarchical decomposition of the whole dataset. This hierarchy can, for example, be constructed by starting with the whole data space as one cluster and then iteratively splitting it into smaller clusters while keeping its relationship to its parents. Alternatively, it can be constructed by considering each data object as a cluster and then iteratively merging those into bigger clusters. The first variant is called the 'top-down approach', also known as 'divisive', and the second variant is called 'bottom-up approach', also known as 'agglomerative'.

We implemented the agglomerative variant of hierarchical clustering. The algorithm decides which two clusters to combined into one by assessing the dissimilarity between the observations in each one using an indicator. Generally, this indicator consists of two components, one is called *metric*, which measures the distance between two observations, and the other is called *linkage*, which measures the degree of dissimilarity between two sets. At each stage of the clustering process, the two clusters that have the smallest linkage distance, are linked together. The choice of metric and linkage has great influence on the final clustering results.

The hierarchical structure can be represented using a *dendrogram*, which is a type of tree diagram showing the relationships between similar sets of data and the distances between dissimilar groups. In a dendrogram, the height of each node is proportional to the dissimilarity between the two child nodes representing clusters being merged. The leaf nodes at the bottom represent a cluster containing only one single data point. At each stage, the two branches with the least distance between them will merge.

Commonly used metrics to measure the distance $d(\boldsymbol{a}, \boldsymbol{b})$ between instances $\boldsymbol{a}$ and $\boldsymbol{b}$ include Euclidean distance, Manhattan distance and Mahalanobis distance.

Some commonly used linkage criteria between two clusters of observations C_i and C_i to measure the distance d_{ij} between them are:

- Complete-linkage clustering: $d_{ij} = \max\{d(\boldsymbol{a}, \boldsymbol{b}) : \boldsymbol{a} \in C_i, \boldsymbol{b} \in C_j\}$
- Single-linkage clustering: $d_{ij} = \min\{d(\boldsymbol{a}, \boldsymbol{b}) : \boldsymbol{a} \in C_i, \boldsymbol{b} \in C_j\}$
- Ward's minimum variance method [14] minimizes the sum of squared differences within all clusters and requires Euclidean distance metrics. It is especially well suited for large datasets. The distance between two clusters determines how much the sum of squares will increase when merged.

  $$d_{ij} = \frac{|C_i| \cdot |C_j|}{|C_i| + |C_j|} ||\boldsymbol{\mu}_i - \boldsymbol{\mu}_j||^2 \tag{10}$$

  where $\boldsymbol{\mu}_i$ is the center of cluster C_i.

The single-linkage clustering method tends to form long thin clusters (chaining phenomenon) because it combines data points linked by a sequence of elements being close to each other. This behaviour is undesirable in our case as the health trend is continuous and a healthy person could end up being joined to a sick person through a series of people with slightly different biomarker values.

The complete linkage is forming compact clusters, but might produce clusters where data points from different clusters are more similar than data points within a cluster.

The Ward's linkage is most well suited for large datasets and achieved the best results in some of the existing research for clustering clinical datasets [7].

4.2 Implementation

We clustered the male and female data separately since they normally have a different range of normal values for most medical features and different risks of diseases.

When performing age prediction we used some calculated metrics as measurement for different models. When using unsupervised learning it can be difficult to compare the different clustering results. Although there exist some commonly used criteria such as *Davies-Bouldin Index* [4] and *Sihouette Coefficient* [11], they were unable to reflect the intrinsic information of a dataset from a medical perspective and hence are not well suited for our task. Our final goal is to find meaningful divisions and to define relevant health classes. Thus, our decision was mainly based on the medical interpretation of the resulting clusters when comparing different clustering results. To achieve this we had to gain a thorough medical understanding of the discovered clusters by conducting a detailed analysis into each feature.

4.2.1 Feature Selection

Most clustering methods cannot handle high-dimensional data efficiently. As we have more than 40 bio-markers, we decided to select only the most relevant features for our clustering using the following methods.

The previously mentioned review score can be used as an indication of one's general health, which is why we selected the features which were most contributing in the MARS review score prediction for clustering. The results of these predictions can be seen in Table 10 and Table 11 in the appendix for males and females respectively.

We started from the features selected from MARS and did further research regarding the common health problems of working people (especially middle-aged elder ones), since we wanted to try to detect risks for some of these diseases in our clustering. These diseases and the major abnormal indicators in health examinations are:

1. Overweight and Obesity: is indicated by high body fat percentage, hip-to-waist ratio, BMI etc. We used body fat percentage, BMI and visceral fat, as they were considering different aspects to indicate obesity problems. For males, a high body fat percentage is more common than in females and seems to be more relevant to their general health. For females, body fat percentage was replaced by body water percentage for clustering as these features are highly correlated in our dataset and the former contains twice as many missing values as the latter.

2. Hyperlipidemia: is an elevation of lipids (fat) in the blood and indicated by high triglycerides, high cholestrol and low high-density lipoprotein (HDL). We selected these three features for clustering. Hyperlipidemia increases the risk of heart disease and stroke.

3. Fatty liver: is indicated by several features such as γ-glutamyltransferase (GGT), transaminase, etc. The fatty liver index (FLI) in our dataset was calculated by an algorithm based on BMI, waist circumference, triglycerides and GGT and has been shown to have an accuracy of 0.84 in detecting fatty liver [2]. FLI varies between 0 and 100. For an FLI below 30 risk of fatty liver disease is extremely low and for an FLI above 60 very high.

4. Osteoporosis: is indicated by a low bone mineral density. In our dataset we only have bone mineral mass and this is not well suited as an indication for osteoporosis. Hence, we had to disregard this disease.

5. Hypertension: is indicated by an elevated blood pressure. We used mean arterial pressure (MAP) on the left arm as a combination of systolic and diastolic blood pressure.

6. Diabetes: is mainly indicated by high fasting glucose level. In our dataset, however, half of the results were taken shortly after the subject had consumed food. The indication of high blood sugar might not be trustworthy.

After several attempts, we noticed that the selection of similar and highly-related features such as waist size, body fat and weight would make the clustering prone to be biased towards specific health conditions such as obesity. As we were keen to consider the general health conditions we decided to also take the correlations between features into consideration and avoided choosing pairs of features which were highly correlated (Pearson correlation coefficients ≥ 0.6)

Furthermore, it was important to take age into account when defining one's health conditions. As we wanted to include features from every part of the body, we decided to also select forced expiratory volume in 1 second (FEV1) for clustering. It was selected due to it being the lung function feature with the highest correlation to the review score.

After trying different feature combinations using the methods mentioned above, the most interpretable clustering results were achieved using the features listed in Table 2.

4.2.2 Pre-processing

Since most clustering methods cannot handle missing values and we wanted to consider as many people as possible for defining their health conditions, we decided to estimate the missing vales and outliers in our dataset as described in Section 2.2. Entries with values of the selected features that were still missing after their estimation were removed.

The clusters are largely based on the distances between the samples within each feature, which are sensitive to differences in magnitude and scales of the attributes. Therefore, dealing with extreme values and scaling the features into a standard range plays a crucial

Feature	male	female	clipped
Fatty liver index	✓	✓	
Mean arterial pressure left	✓	✓	
Body fat percentage	✓		
Body water percentage		✓	
Body mass index	✓	✓	
Triglycerides	✓	✓	✓
Cholesterol	✓	✓	✓
High density lipoprotein	✓	✓	✓
Visceral fat level	✓	✓	
Age	✓	✓	
Forced expiratory volumn in 1 second	✓	✓	✓

Table 2: Features selected for clustering

role in successful clustering. The goal is to equalize the size and variability of these features.

For some features which contained very few extreme values such as triglycerides, cholesterol and HDL, we decided to clip the 0.5% and 99.5% quantiles. Doing so will not lead to the loss of much information as the values which are clipped will still be very extreme and hence indicate a severe illness. The advantage of the clipping is that we achieve more similar ranges after scaling all the features. The list of clipped features can be found in Table 2.

Commom methods for feature scaling include:

- Standardization:

$$x_i' = \frac{x_i - \mu(x)}{\sigma(x)} \tag{11}$$

  with μ being the sample mean and σ the standard deviation. Standardization assumes the data is (nearly) normally distributed within each feature and will scale them such that the distribution is centred around 0, with a standard deviation of 1.

- Min-max scaling:

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{12}$$

  It rescales the range of feature in $[0, 1]$ by using a linear mapping.

Min-max scaling is more suited for uniformly distributed data and since outliers (mildly extreme values) in our dataset have their medical meaning, we wanted to leave them far enough from normal healthy values by standardization and to not shrinks to $[0, 1]$ as the min-max scaler does. We therefore decided to use standardization to scale our features.

4.2.3 Visualization

To evaluate the performance of our clustering, we decided to use some visualization tools to help us understand the clusters we got. Since we were dealing with high dimensional data, which was hard to visualize, some dimensionality reduction techniques need to be introduced.

The *t-distributed stochastic neighbor embedding* (t-SNE) [10] is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions. Specifically, it models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability.

First, t-SNE constructs a probability distribution over pairs of objects in such a way that similar objects are assigned a higher probability while dissimilar points are assigned a very low probability. Second, t-SNE defines a similar probability distribution over the points in the low-dimensional map, and it minimizes the Kullback-Leibler divergence (KL divergence) between the two distributions with respect to the locations of the points in the map.

As a result, t-SNE allowed us to visualize the distribution of all data points in a low-dimensional space and helped us to better explore the clusters we found.

5 Results and Discussions

5.1 Age Prediction

We started by predicting the age using MLR, the state of the art as described in section 1.2. The features with the highest correlation to age were selected for the prediction. Though we made sure to check the correlation between all the selected features and remove one of two highly correlated ones. The feature selection can be found with coefficients in appendix B.1. We decided to use the cleaned data before estimating missing values since some of our estimations were calculated based on age and could introduce bias in age prediction. Rows with a NaN value within one of those features were dropped. The resulting prediction had a mean absolute error (MAE) of 7.84, which is higher than expected, as even when predicting the average age one can already achieve a MAE of 9.36. We wanted to see if we can improve the MAE with a more complex model. Therefore, we tried Lasso, Random Forest, MARS and Neural Networks with different methods of feature selection. All of those models were approximately in the same range with a MAE around 7.3.

The best result was achieved by a GAM with a slightly lower MAE of 7.21. Features were selected by iteratively eliminating the features with the least feature importance. The final age prediction was done using 17 selected features of high feature importance which are listed in appendix B.2. Figure 2 shows two figure functions of the trained model.

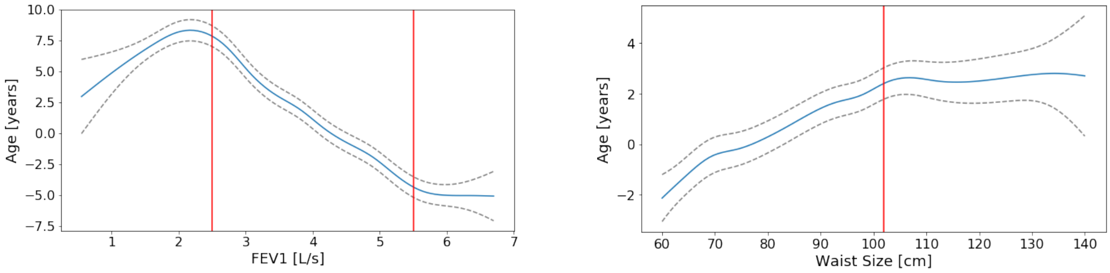


Figure 2: The left plot shows the feature function of the forced expiratory volume in 1 second of the age prediction model with two red lines indicating the normal range. The right plot shows the feature function of the waist size of the age prediction model. Everything on the right side of the red line is considered abnormal. The dashed lines are the confidence intervals.

The feature function from the forced expiratory Volume in 1 second (FEV1) clearly shows nonlinear dependencies outside of the normal range, indicated by the two red lines. This can be explained by the fact that a value below 2.5 is a likely indication of an illness. Consequently, the lowest values are not age-dependent which is also shown by the broader confidence interval on the left side, indicated by the dashed lines. Similarly, in the feature function of the waist size, we have a sudden change in slope approximately at the point where the waist size cannot be considered normal anymore (indicated by the red line). For comparison, we trained a MLR model on the same set of features and achieved a MAE of 7.29, which is only slightly worse than the MAE achieved by the GAM. This is likely due to the fact, that most of the data is in the normal range where we find a linear trend, which can be predicted just as well by the MLR model. Based on our results we can establish that a MLR model can perform very well when all values are in a normal range. Nonetheless, we do have abnormal values in our data and therefore it does makes sense to stick with a more complex model like GAM, to achieve a better accuracy.

Nevertheless, we want a lower age to be correlated with good biomarkers and a higher age with worse biomarkers. This is achieved by just keeping the linear part of the functions shown in figure 2, which can be represented by a MLR model. Unfortunately, not all the feature functions of the GAM have linear dependencies in their normal range meaning that they cannot be accurately predicted by MLR which is why we predict the age using both MLR and GAM.

Not only does the age prediction allow us to calculate a potential indicator of general health, but it also gives us the chance to explore the limits and potentials of our synthetic dataset. We asked Wellabe to train a GAM on the original data to allow us to compare its performance to ours. The result is shown in figure 3. On the left, we can see the outcome of the model trained on the original data with a MAE of 2.3. The right plot shows the outcome of their model trained on the synthetic data. It is worth noting that the synthetic data used here has not been pre-processed. The MAE is 6.82.

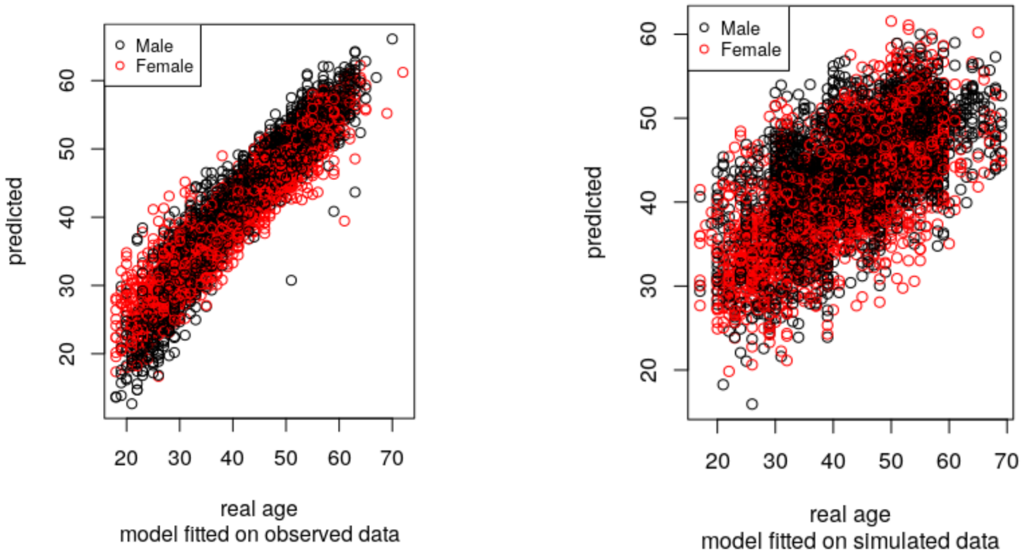


Figure 3: The left plot shows the predicted age over real age of a GAM trained and tested on the observed/original data. On the right side is the same plot of a model trained and tested on the simulated/synthetic data.

The MAE for the synthetic data is 0.4 better than what we were able to achieve. This could be due to the fact that we tried to minimize the number of features used, as having too many features could lead to overfitting and might introduce noise. Another reason could be our pre-processing methods which could have lead to a loss of information. Still, their result is only slightly better than ours which confirms that there is not much more information to be gained in age prediction. Additionally, these plots clearly show that there are limits when using synthetically generated data when predicting a single instance.

Coming back to the original intention of gaining an indicator of general health. The age difference is added as an additional feature. Unfortunately, we could not find significant and meaningful patterns in comparison to the other features. This probably comes from the bad approximation of the age itself, as not all the information according to age could be transferred from the real data to the synthetic one.

5.2 Review Score Prediction

When predicting the review score both GAM and MARS were able to perform well. Their results are shown in table 3. There is a good chance that the review scores are subjective as they were not always given by the same doctor. We thus did not expect to get perfect prediction results and it is, therefore, reasonable to consider the resulting MAE's around 0.8 (for a score between 0 and 6) as a good performance.

	Baseline	MARS	GAM
MAE	1.21	0.82	0.85

Table 3: Review score prediction: mean absolute error

The prediction using MARS gave us a list of features which are strongly related to the review score such as fatty liver index and blood pressure readings. The full list can be found in the Appendix, tables 10, 11. As the review score is one of the indicators of general health, we considered these features as most relevant in determining one's overall health. Therefore, we used this list of features as a base for feature selection for clustering.

5.3 Clustering

Although K-means is a popular choice for clustering due to its simplicity and efficiency, for our dataset it was unable to find any medically meaningful clusters. Regardless the number of clusters chosen, all resulting partitions were of similar size and could not be interpreted well.

The advantage of DBSCAN is that it does not require you to specify the number of clusters in advance. However, in our case, DBSCAN was unable to ever detect more than one significant cluster. Tuning the hyper-parameters just altered the amount of data labelled as noise. A possible reason for this is the fact that health is continuous and not categorical making it very hard to detect distinct clusters.

The hierarchical clustering forces the formation of clusters through a hierarchical structure, which provides the possibility to divide each cluster into increasingly detailed smaller clusters. As we expected, complete and single-linkage functions did not provide reasonable clusters on our datasat. Using Ward linkage (which forced us to use Euclidean distance metric) produced the most interpretive clusters among all the linkage functions. We used t-SNE embeddings to visualize our clustering results. We analysed these visualizations with different numbers of clusters selected. By investigating them and the dendrogram (Figure 4) of the hierarchical clustering, we chose the number of clusters for females and males as 5 and 6 respectively and these most medically relevant clusters can be seen in Figure 5, where the clusters were mostly clearly separated. As expected, some small clusters with unusual health conditions are found.

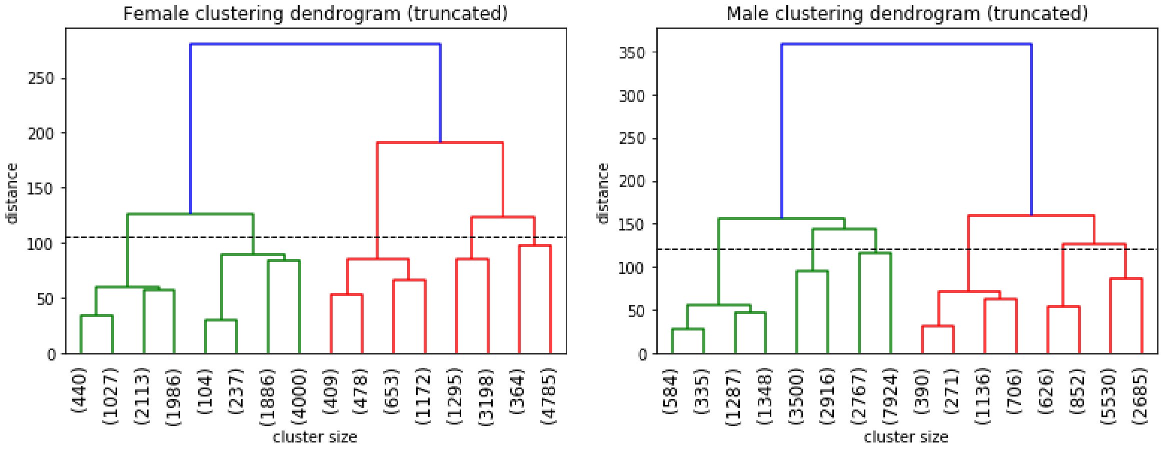


Figure 4: Truncated dendrogram for hierarchical clustering on females (left) and males (right), 10 features selected, horizontal line indicates the distance threshold of our selection of number of clusters

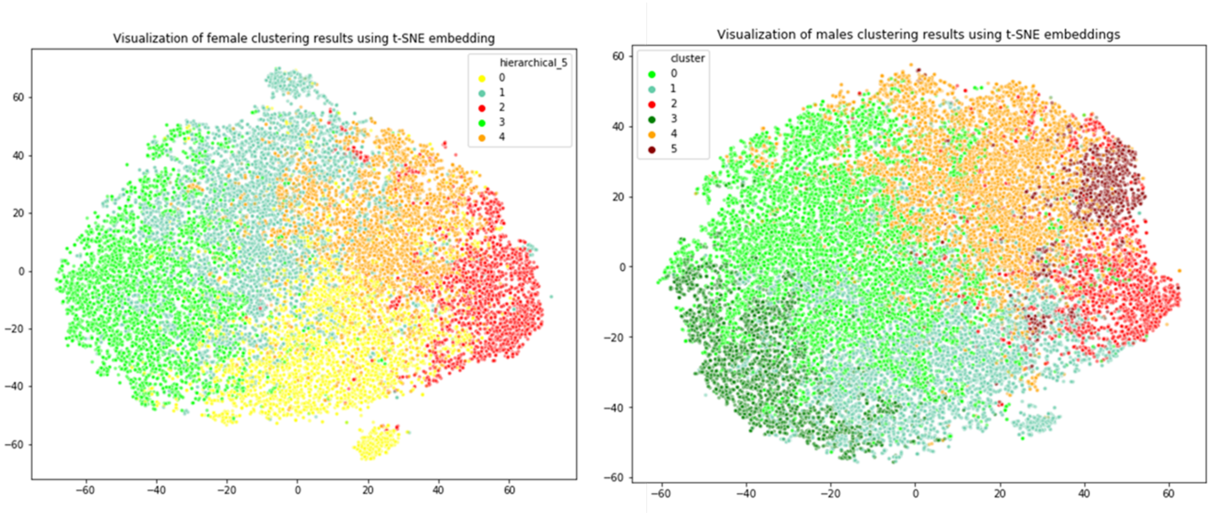


Figure 5: Clustering results of females (left) and males (right) using t-SNE embeddings

We then explored the clusters for females and males in details and both of them gave meaningful results which were also mostly consistent with the review scores given by health experts.

5.3.1 Females

The most medically relevant clusters were found when performing hierarchical clustering with 5 clusters. The number stated in the bracket is the size of the group.

- Healthy people, marked in green in the t-SNE (11793, 49%)
  - Cluster 1 (6227, 26%): middle aged, most values are in the normal ranges;
  - Cluster 3 (5566, 23%): young, low fatty liver index, low mean arterial pressure, low visceral fat, high forced expiratory volume in 1 second and high body water;
- People with elevated risk, marked in yellow, orange and red in the t-SNE (12354, 51%)
  - Cluster 0 (5149, 21%): middle aged, lower body water and higher body fat, all other values are normal, but slightly worse than in cluster 1 and cluster 3;
  - Cluster 2 (2712, 11%): high fatty liver index, high body mass index, high triglycerides, high mean arterial pressure and high visceral fat;
  - Cluster 4 (4493, 19%): older people, elevated mean arterial pressure, lower body water

Overall, cluster 3 is the healthiest group with the best checkup results and mostly perfect review scores of '6'. Cluster 1 is a cluster of average healthy people, their results are normal.

Cluster 0 has slightly worse results, but they are not yet of concern. This group is characterized by an elevated risk of being overweight. Cluster 4 is the group containing people with hypertension problems leading to an elevated cardiovascular diseases risk. Finally, cluster 2 contains unhealthy people with already existing health problems. Almost all people with a review score of '1' are assigned to this group.

5.3.2 Males

After investigating the dendrogram (Figure 4) and gaining a medical understanding of each cluster, the number of clusters was set to 6 for males. All six clusters have a unique medical interpretation. It is worth noting that three of them can be combined to build a larger generally low-risk group and the other three a larger generally high-risk group. The difference on some selected features between these two larger groups could also be spotted in the boxplots in the appendix (Figure 7). In general, almost two thirds (63%) of all males are considered to be low-risk and healthy. Most of them are young with a good lung function, low blood pressure and a low risk of fatty liver disease and obesity. While most of the people in the high-risk group have a fatty liver index (FLI) larger than 60 (high risk of fatty liver disease), most low-risk people have an FLI below 60.

A more detailed description of each group can be found below. The visualization of the comparison between the groups based on some key features can further be found in the appendix (Figure 8 and 9). Altogether, we had 32857 male instances.

- People with low health risks, marked in green in the t-SNE plot (20661, 63%)
  - Cluster 3: (3554, 11%) young, very low FLI, visceral fat and body fat percentage, high forced expiratory volume in 1 sec (FEV1).

  - Cluster 1: (6416, 19%) middle-aged, slightly high body fat percentage, FLI larger than 30 and mostly less than 60, high FEV1.
  - Cluster 0: (10691, 33%) old, slightly high blood pressure, slightly higher FLI than cluster 3 but lower than cluster 1.

- People with high health risks, marked in orange and red in t-SNE (12196, 37%)
  - Cluster 4: (8215, 25%) middle-aged, very high triglycerides and low high-density lipoprotein
  - Cluster 5: (1478, 4%) middle-aged, very high fatty liver index (above 80), high body fat percentage and visceral fat
  - Cluster 2: (2503, 8%) old, high fatty liver index but lower than cluster 5

Overall, cluster 3 is the healthiest group containing young and fit people who also received the best review scores from the doctor's. Cluster 1 can be considered being between low and medium risk since it has the highest risk of obesity and fatty liver disease among the low-risk groups. Cluster 0 is the largest group and contains on average older people than the other 2 low-risk groups but they are in mostly good health with potential risk for fatty liver disease and hypertension. People in the latter two groups received medium to high scores from the doctor's. They should do more regular health checkups, take better care of their nutrition and increase their amount of physical exercise, whereas cluster 3 should stick with their healthy habits.

The high risks groups are more crucial to investigate in details. All three groups contain people with varying degrees of hypertension.

Cluster 4 contains almost 25% of all males and has an extremely high triglycerides and low HDL. Hence people in this group show clear signs of hyperlipidemia (high blood fat level) which increases the risk of heart disease and stroke. Their fatty liver index is also mostly above 60. Although they are mostly just middle-aged, they already are at very high risk of disease.

Cluster 5 is the smallest group and with the highest risk of fatty liver disease and obesity. People in this group should change their diet and limit their alcohol consumption. Cluster 2 contains old people with a high risk of fatty liver disease. Though their risk is not as high as the ones of people in the other two groups. They also mainly received low to medium review scores from the doctors, whereas the other two groups mainly received negative evaluations.

People in all three high risk groups should consult a doctor and change their diet and exercise routines.

6 Conclusion and Outlook

By exploring our dataset in great detail and implementing various models we were able to achieve our main goal and successfully define meaningful health classes for males and females. The results stayed consistent with the review scores given by health experts and successfully detected high risk groups for some common diseases. As our resulting classes are not only able to identify people with very good and average health but also separate people with different diseases, they also pose great potential in helping Wellabe improve their recommendations of prevention plans and programs for each individual.

We successfully implemented various algorithms to explore our first side objective of age prediction. Comparing the prediction performance of our data with the one of the original data gave us some insights into the limitations of synthetic data. This was also part of our second side objective of identifying the limitations and potentials of synthetically generated data. This objective was successfully accomplished as we did not only find some of its limitations but also showed its potential by revealing that the common patterns replicated in the synthetic data were sufficient enough for the successful clustering of medically relevant groups.

Overall we were able to make a valuable contribution to Wellabe with meaningful health classes based on a hierarchal structure which also presents the potential for further division and exploration. Furthermore, our positive results encourage further investigation into clustering of synthetically generated health data.

Acknowledgements

We would like to express our sincere appreciation to Dr. Dominik Jüstel and Jan Kukačka from Helmholtz Zentrum for their valuable and constructive feedback and suggestions during the planning and execution of this project. We would also like to thank Dr. Sebastian Dünnebeil and Ervin Hasko from Wellabe for allowing us to use their data and constantly being willing and able to assist us in achieving the best possible outcome of this project. We also want to thank our co-mentor Stefan Bamberger, who provided us valuable feedback throughout the project. Last but not the least, we want to thank Prof. Dr. Massimo Fornasier and Dr. Ricardo Acevedo Cabra from TUM for organizing the Data Innovation Lab for several years and providing us with the opportunity of dealing with real data science problems.

References

[1] *A Tour of pyGAM.* https://pygam.readthedocs.io/en/latest/notebooks/tour_of_pygam.html. Accessed: 6 July 2020.

[2] Giorgio Bedogni et al. "The Fatty Liver Index: A simple and accurate predictor of hepatic steatosis in the general population". In: *BMC gastroenterology* 6 (Nov. 2006), p. 33. DOI: 10.1186/1471-230X-6-33.

[3] Wojtek J. Chodzko-Zajko. "The physiology of aging: Structural changes and functional consequences. Implications for research and clinical practice in the exercise and activity sciences". English (US). In: *Quest* 48.3 (Aug. 1996), pp. 311–329. ISSN: 0033-6297. DOI: 10.1080/00336297.1996.10484200.

[4] D. L. Davies and D. W. Bouldin. "A Cluster Separation Measure". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2 (1979), pp. 224–227.

[5] T. Furukawa et al. "Assessment of Biological Age by Multiple Regression Analysis". In: *Journal of Gerontology* 30.4 (2012), pp. 422–434.

[6] Fedan KB Hankinson JL Odencrantz JR. "Spirometric reference values from a sample of the general U.S. population". In: *Am J Respir Crit Care Med* (1999).

[7] Shoji Hirano, Xiaoguang Sun, and Shusaku Tsumoto. "Comparison of clustering methods for clinical databases". In: *Information Sciences* 159.3 (2004), pp. 155 – 165. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2003.03.011. URL: http://www.sciencedirect.com/science/article/pii/S0020025503001956.

[8] M. Liao et al. "Cluster analysis and its application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis". In: 17 (2016), p. 25.

[9] B. Loong, A. M. Zaslavsky, and Y. He. "Disclosure Control using Partially Synthetic Data for Large-Scale Health Surveys, with Applications to CanCORS". In: 32 (2013), pp. 4139–4161.

[10] Laurens van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE". In: *Journal of machine learning research* 9.Nov (2008), pp. 2579–2605.

[11] Peter J. Rousseeuw. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53 –65. ISSN: 0377-0427. DOI: https://doi.org/10.1016/0377-0427(87)90125-7. URL: http://www.sciencedirect.com/science/article/pii/0377042787901257.

[12] E. D. Ubeyli and E. Dodgu. "Automatic Detection of Erythemato-Squamous Diseases Using k-Means Clustering". In: *Journal of Medical Systems* 34 (2010), pp. 179–184.

[13] L.M. Ueno, Y. Yamashita, and T. Moritani. "Biomarkers of Aging in Women and the Rate of Longitudinal Changes". In: *Journal of PHYSIOLOGICAL ANTHROPOLOGY and Applied Human Science* 22 (2003), pp. 37–46.

[14] Joe H Ward Jr. “Hierarchical grouping to optimize an objective function”. In: *Journal of the American statistical association* 58.301 (1963), pp. 236–244.

[15] P. E. Watson, I. D. Watson, and R. D. Batt. “Total body water volumes for adult males and females estimated from simple anthropometric measurements”. In: *The American journal of clinical nutrition* 33 (1980), pp. 27–39.

[16] Ming-Ni Wu, Chia-Chen Lin, and Chin-Chen Chang. “Brain Tumor Detection Using Color-Based K-Means Clustering Segmentation”. In: vol. 2. Dec. 2007, pp. 245–250. ISBN: 978-0-7695-2994-1. DOI: 10.1109/IIHMSP.2007.4457697.

Appendices

A Data Understanding and Handling

Feature Name	Abbreviation	Normal Range	Description
Sex	sex	1 for male, 2 for female	—
Age	age	18 - 65	—
Review score	review_score	integer between 0 - 6	A score given by a doctor as a judgement of health status. Score of 0 means medical emergency and 6 means optimal health.
Systolic Blood Pressure	SYS_BP_RIGHT, SYS_BP_LEFT	90 - 120 mmHg	Amount of pressure in the arteries during the contraction of the heart muscle
Diastolic Blood Pressure	DIA_BP_RIGHT, DIA_BP_LEFT	60 - 80 mmHg	Blood pressure when the heart muscle is between beats
Pulse Pressure	PP_RIGHT, PP_LEFT	40 - 60 mmHg	The difference between systolic and diastolic blood pressure
Mean Arterial Pressure	MAP_RIGHT, MAP_LEFT	70 - 100 mmHg	The average of blood pressure over a cardiac cycle
Ankle-Brachial Index	ABI_L, ABI_R	0.9 - 1.29	The ratio of the blood pressure at the ankle to the blood pressure in the upper arm (brachium)
Resting Heart Rate	HEART_RATE	60 - 100 bpm	The number of times the heart beats per minute when it is at rest
Blood Sugar	GLUCOSE_LEVEL	70 - 99 $\frac{\text{mg}}{\text{dl}}$	Concentration of glucose present in the blood of humans
Triglycerides	TG	$< 150\frac{\text{mg}}{\text{dl}}$	Type of fat found in your blood
Total Cholesterol	Cholesterol	$< 200\frac{\text{mg}}{\text{dl}}$	Total amount of cholesterol in your blood (LDL + HDL)
High-density lipoproteins	HDL	$> 40\frac{\text{mg}}{\text{dl}}$	Good cholesterol
Alanine transaminase	ALT	7 - 55 U/L	An enzyme in the liver, converts proteins into energy for the liver cells
Aspartate transaminase	AST	8 - 48 U/L	An enzyme helps metabolize amino acids
Gamma-glutamyl transferase	GGT	8 - 61 U/L	An enzyme that is found in many organs throughout the body, with the highest concentrations found in the liver
Creatine	CREA	0.5 - 1.2 $\frac{\text{mg}}{\text{dl}}$	Indicates amount of muscle a person has and an indication of kidney function
Fat liver index	FLI	< 30	Can detect excess fat in the liver

Table 4: List of features from cardiovascular system and metabolism sections with their descriptions

Feature Name	Abbreviation	Normal Range	Description
Oxygen Saturation	OXYGEN _SATURATION	0.95-0.99	Fraction of oxygen-saturated hemoglobin relative to total hemoglobin in the blood.
Forced Vital Capacity	FVC_L	2.5 - 4.5 liters	The maximum amount of air that can be exhaled from your lungs after breathing in as much air as possible.
Forced Expiratory Volume in 1 Second	FEV1_L	0.7-1 *FVC_L	The maximum air one can breath out in 1 second.
Peak Expiratory Flow	PEF_L	5-9.5 $\frac{liters}{s}$	The maximum speed someone can breath out.
Forced Expiratory Flow at 25% of the Lung Volume	FEF25	1.5-5 $\frac{liters}{s}$	The speed of the air flow when there is only 25% air of the lung volume left.
Body Height	HEIGHT	BMI 18-30	Body height
Body Weight	WEIGHT	BMI 18-30	Body weight
Visceral Fat Level	VISCERAL_FAT	≤ 9	Fat stored with abdominal cavity and not visible from the outside
Daily Caloric Needs	DAILY_CALORIC _NEEDS	1600-2400 kcal female; 2000-3000 kcal male	Number of calories a person needs to remain healthy
Muscle Mass	MUSCLE_MASS	24.2-30.3% of weight female; 33.2-39.4% of weight male	Weight of muscles
Body Fat Mass	FAT_KG	normal range for fat % times weight	Weight of body fat
Body Fat Percentage	BODY_FAT	14-32% female; 5-25% male	Body fat mass in percentage of body weight
Body Water Percentage	BODY_WATER	41-60% female; 43-73% male	Total water in body as a percentage of body weight
Bone Mineral Mass	BONE_MINERAL _MASS	1.8-3.35 kg female; 2.45-4.1 kg male	Amount of bone mineral in bone tissue
Hip Size	HIP_SIZE	80-120cm	Circumference of the upper swell of the hip
Waist Size	WAIST_SIZE	60-120cm	Circumference of the waist

Table 5: List of features from respiratory system and body composition sections with their descriptions

Feature name	Outlier detection	Estimation of missing values
Systolic Blood Pressure	(SYS_BP<40) or ((SYS_BP<90) and (SYS_BP < DIA_BP)	measurement on other side, or PP+DIA_BP, or $3 \cdot$ MAP $-$ $2 \cdot$ DIA_BP
Diastolic Blood Pressure	(DIA_BP< 30) or ((DIA_BP < 40) and ($\lvert$DIA_BP_LEFT $-$ DIA_BP_RIGHT$\rvert$ > 30))	measurement on other side, or SYS_BP - PP, or $3 \cdot$ MAP $-$ SYS_BP
Pulse Pressure	(PP < 10) or ((PP $\notin [20, 90]$) and ($\lvert$(SYS_BP $-$ DIA_BP) $-$ PP$\rvert$ > 30))	measurement on other side, or SYS_BP - DIA_BP
Mean Arterial Pressure	(MAP <20) or ((MAP $\notin [70, 140]$) and not(DIA_BP < MAP < SYS_BP))	measurement on the other side, or $2/3 \cdot$ DIA_BP $+ 1/3 \cdot$ SBP_BP
Ankle-Brachial Index	(ABI< 0.3) or ((ABI < 0.5) and ($\lvert$ABI_L$-$ABI_R$\rvert$ > 0.45))	measurement on the other side
Resting Heart Rate	(HEART_RATE < 30)	—
Blood Sugar	GLUCOSE_LEVEL < 60	Median of each age group
Triglycerides	(TG == 0) or (CHOLESTEROL <= HDL + $0.2 \cdot$ TG)	Median of each age group
Total Cholesterol	(Cholesterol == 0) or (CHOLESTEROL <= HDL + $0.2 \cdot$ TG)	Median of each age group
High-density lipoproteins	(HDL == 0) or (CHOLESTEROL <= HDL + $0.2 \cdot$ TG)	Median of each age group
Alanine transaminase	ALT == 0	Median of each age group
Aspartate transaminase	AST == 0	Median of each age group
Gamma-glutamyl transferase	GGT == 0	Median of each age group
Creatine	CREA == 0	Median of each age group
Fat liver index	(FLI< 0) or (FLI > 100)	$\frac{e^{0.9 \cdot \log(\mathrm{TG}) + 0.1 \cdot (\mathrm{BMI}) + 0.7 \cdot \log(\mathrm{GGT}) + 0.05 \cdot (\mathrm{WC}) - 15.7}}{1 + e^{0.9 \cdot \log(\mathrm{TG}) + 0.1 \cdot (\mathrm{BMI}) + 0.7 \cdot \log(\mathrm{GGT}) + 0.05 \cdot (\mathrm{WC}) - 15.7}} \cdot 100$

Table 6: Data Cleaning for cardiovascular system and metabolism

Feature name	Outlier detection	Estimation of missing values
Oxygen Saturation	< 60	Mean of female or male patients
Forced Vital Capacity	< 1 or FEV1_L $>$ FVC_L	male: $-0.1933+0.00064\cdot \text{age}-0.00269\cdot \text{age}^2+0.00018642\cdot \text{height}^2$ female: $-0.3560+0.01870\cdot \text{age}-0.000382\cdot \text{age}^2+0.00014815\cdot \text{height}^2$
Forced Expiratory Volume in 1 Second	< 0.3	male: $0.5536-0.01303\cdot \text{age}-0.000172\cdot \text{age}^2+0.00014098\cdot \text{height}^2$ female: $0.4333-0.00361\cdot \text{age}-0.000194\cdot \text{age}^2+0.00011496\cdot \text{height}^2$
Peak Expiratory Flow	< 2	male: $1.0523+0.08272\cdot \text{age}-0.001301\cdot \text{age}^2+0.00024962\cdot \text{height}^2$ female: $0.9267+0.06929\cdot \text{age}-0.001031\cdot \text{age}^2+0.00018623\cdot \text{height}^2$
Forced Expiratory Flow at 25% of the Lung Volume	PEF_L $<$ FEF25	male: $2.7006-0.04995\cdot \text{age}+0.00010345\cdot \text{height}^2$ female: $2.3670-0.01904\cdot \text{age}-0.000200\cdot \text{age}^2+0.00006982\cdot \text{height}^2$ [6]
Body Weight	calculated BMI< 10 or > 60	$(\text{BODY\_FAT}-(0.16\cdot \text{age})-(10.34\cdot(\text{sex}-2))+9)/1.39)\cdot \text{height}^2/10000$
Visceral Fat	—	—
Daily Caloric Needs	< 200 or $< 60\%$ of calories burned at rest $(9.99\cdot \text{weight}+6.25\cdot \text{height}-4.92\cdot \text{age}-200)$	$655.1+(4.35\cdot \text{weight})+(4.7\cdot \text{height})-(4.7\cdot \text{age})$ female;$66+(6.2\cdot \text{weight})+(12.7\cdot \text{height})-(6.76\cdot \text{age})$ male
Muscle Mass	< 10	$0.252\cdot \text{weight}+0.473\cdot \text{height}-48.3$ female; $0.407\cdot \text{weight}+0.267\cdot \text{height}-19.2$ male
Body Fat Mass	==0 or calculated weight from BODY_FAT and FAT_KG invalid but BODY_FAT normal	$(\text{BODY\_FAT}\cdot \text{weight}/100)$
Body Fat Percentage	$< 7\%$ female; $< 3\%$ male or calculated weight from BODY_FAT and FAT_KG invalid but FAT_KG normal	$\text{weight}\cdot 1.39\cdot 10000/\text{height}^2+0.16\cdot \text{age}+10.34\cdot(\text{sex}-2)-9$
Body Water Percentage	$< 35\%$ female; $< 40\%$ male	$(2-\text{sex})\cdot(2.447-0.09156\cdot \text{age}+0.1074\cdot \text{height}+0.3362\cdot \text{weight})+(\text{sex}-1)\cdot(-2.097+0.1069\cdot \text{height}+0.2466\cdot \text{weight}))/\text{weight}\cdot 100$
Bone Mineral Mass	< 35	use linear regression from age, sex and weight
Hip Size	< 60	—
Waist Size	< 60	—

Table 7: Data Cleaning for respiratory system and body composition

B Age Prediction

B.1 MLR

Predictor	Coefficient
Intercept	21.711830401357346
FEV1_L	-2.85631147
VISCERAL_FAT	1.12680704
MAP_RIGHT	0.1801872
CHOLESTEROL	0.03450363

Table 8: MLR with a basic feature selection

Predictor	Coefficient
Intercept	35.726177894608426
ABI_L	6.85994154
ABI_R	11.46340514
ALT	-0.03885476
CHOLESTEROL	0.0313574
CREA	2.66621868
DIA_BP_RIGHT	0.16108864
FEF25	-0.86779391
FEV1_L	-3.97077775
GLUCOSE_LEVEL	0.09845636
HDL	0.04839736
HEART_RATE	-0.0486207
HIP_SIZE	-0.18242508
OXYGEN_SATURATION	-0.3251212
PEF_L	0.84430203
PP_RIGHT	0.06757111
VISCERAL_FAT	0.95557046
WAIST_SIZE	0.10798836

Table 9: MLR with GAM feature selection

B.2 GAM Feature Selection

ABI_L, ABI_R, ALT, CHOLESTEROL, CREA, DIA_BP_RIGHT, FEF25, FEV1_L, GLUCOSE_LEVEL, HDL, HEART_RATE, HIP_SIZE, OXYGEN_SATURATION, PEF_L, PP_RIGHT, VISCERAL_FAT, WAIST_SIZE

C Review Score Prediction

Predictor	Coefficient
Intercept	3.3371169
h(145-AST)	0.0037827
h(CHOLESTEROL-139)	-0.0048509
h(74-DIA_BP_LEFT)	-0.0454621
h(DIA_BP_LEFT-74)	-0.0180478
h(21.66-FAT_KG)	0.0451657
h(44-FLI)	0.0225573
h(FLI-44)	-0.0045513
h(21-GGT)	0.0152315
h(129-GLUCOSE_LEVEL)	0.0147904
h(56-HDL)	-0.0189998
h(HDL-56)	0.0079301
h(MAP_LEFT-94)	-0.0208412
h(PP_LEFT-43)	-0.0082112
h(SYS_BP_RIGHT-114)	-0.0068087
h(7-VISCERAL_FAT)	0.0754450
h(BODY_FAT-18) · h(44-FLI)	-0.0010916
h(CHOLESTEROL-217) · h(DIA_BP_LEFT-74)	0.0002756
h(DIA_BP_LEFT-74) · h(38-HDL)	0.0015744
h(74-DIA_BP_LEFT) · h(102.8-WEIGHT)	0.0021784
h(21.66-FAT_KG) · h(MAP_RIGHT-94)	-0.0022905
h(21.66-FAT_KG) · h(94-MAP_RIGHT)	-0.0021000
h(5.35-FEV1_L) · h(HDL-56)	-0.0109946
h(44-FLI) · h(VISCERAL_FAT-9)	-0.0040718
h(44-FLI) · h(9-VISCERAL_FAT)	-0.0015212
h(129-GLUCOSE_LEVEL) · h(110-HIP_SIZE)	0.0004268
h(129-GLUCOSE_LEVEL) · h(109-PEF_%)	-0.0002496
h(SYS_BP_RIGHT-114) · h(111-TRIGLYCERIDES)	0.0002073

Table 10: Review score MARS prediction for males, selected predictors

Predictor	Coefficients
Intercept	3.3829668
h(BODY_FAT-20)	-0.0118968
h(CHOLESTEROL-166)	-0.0115236
h(CHOLESTEROL-221)	0.0085359
h(27.73-FAT_KG)	0.0592986
h(24-FLI)	0.0326769
h(139-GLUCOSE_LEVEL)	0.0113015
h(MAP_LEFT-99)	-0.0253821
h(94-MAP_RIGHT)	-0.0067675
h(MAP_RIGHT-94)	-0.0231941
h(65-PEF_%)	0.0127050
h(PEF_%-65)	0.0080321
h(40-AST) · h(CHOLESTEROL-166)	0.0002278
h(CHOLESTEROL-166) · h(99-HIP_SIZE)	0.0004008
h(27.73-FAT_KG) · h(FLI-24)	-0.0017146
h(16.97-FAT_KG) · h(139-GLUCOSE_LEVEL)	-0.0009393
h(27.73-FAT_KG) · h(MAP_RIGHT-128)	0.0177788
h(27.73-FAT_KG) · h(SYS_BP_LEFT-131)	-0.0016107
h(24-FLI) · h(7.86-PEF_L)	-0.0045162
h(24-FLI) · h(TRIGLYCERIDES-82)	-0.0002465
h(22-GGT) · h(139-GLUCOSE_LEVEL)	0.0002976
h(139-GLUCOSE_LEVEL) · h(VISCERAL_FAT-2)	-0.0007195
h(MAP_RIGHT-94) · h(33.5-MUSCLE_MASS)	-0.0155704
h(PEF_.-65) · h(PP_LEFT-44)	-0.0003696

Table 11: Review score MARS prediction for females, selected predictors

D Clustering

Figure 6: Box plots for hierarchical clustering on females

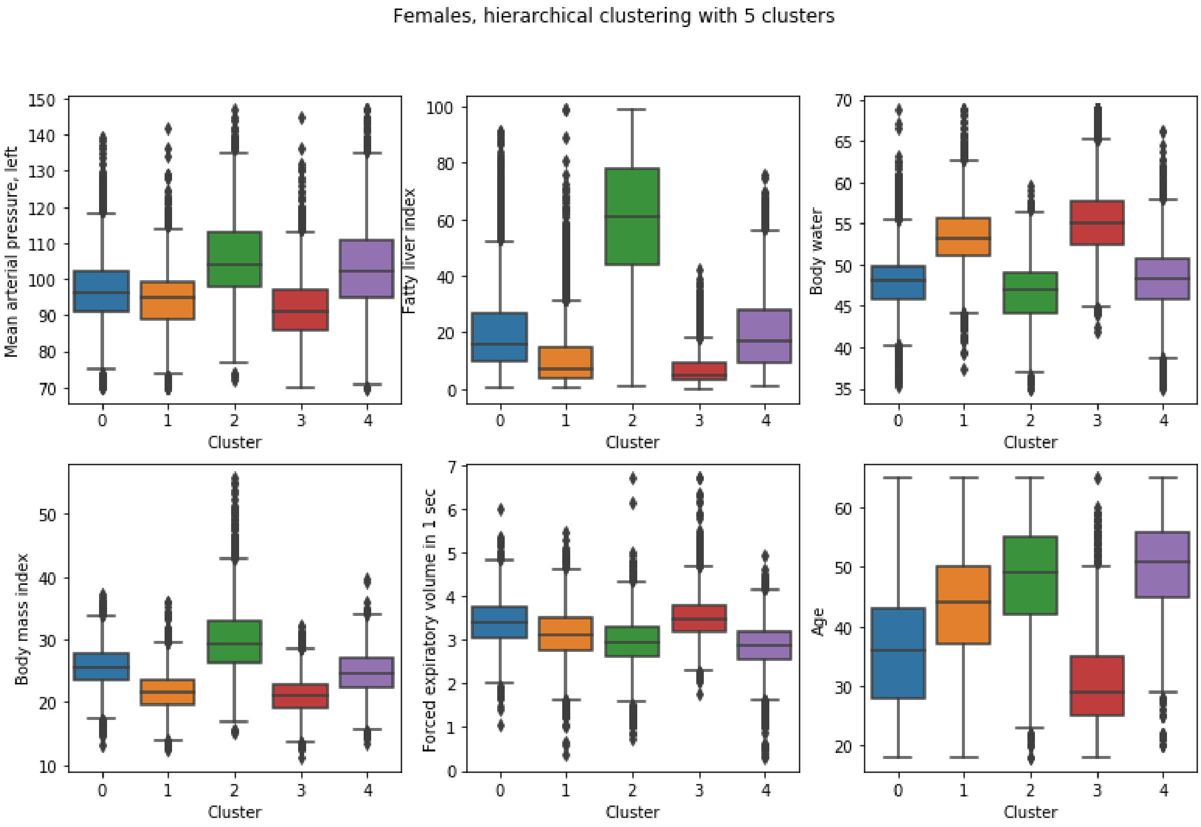


Figure 7: Box plots for high and low risk groups on males

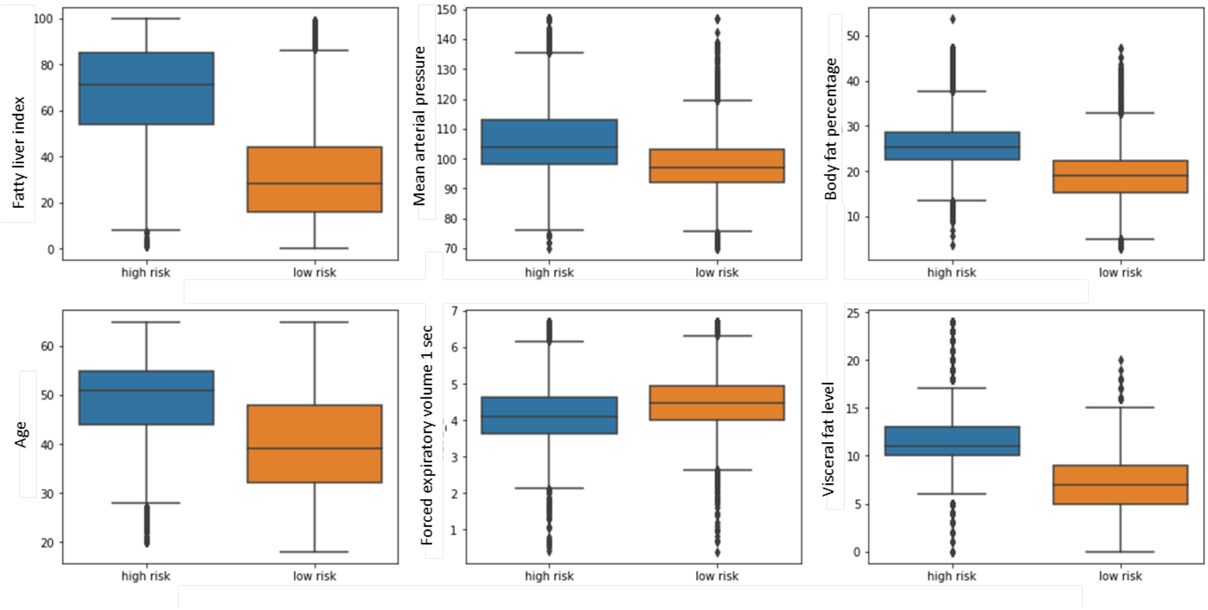


Figure 8: Box plots for low risk groups on males

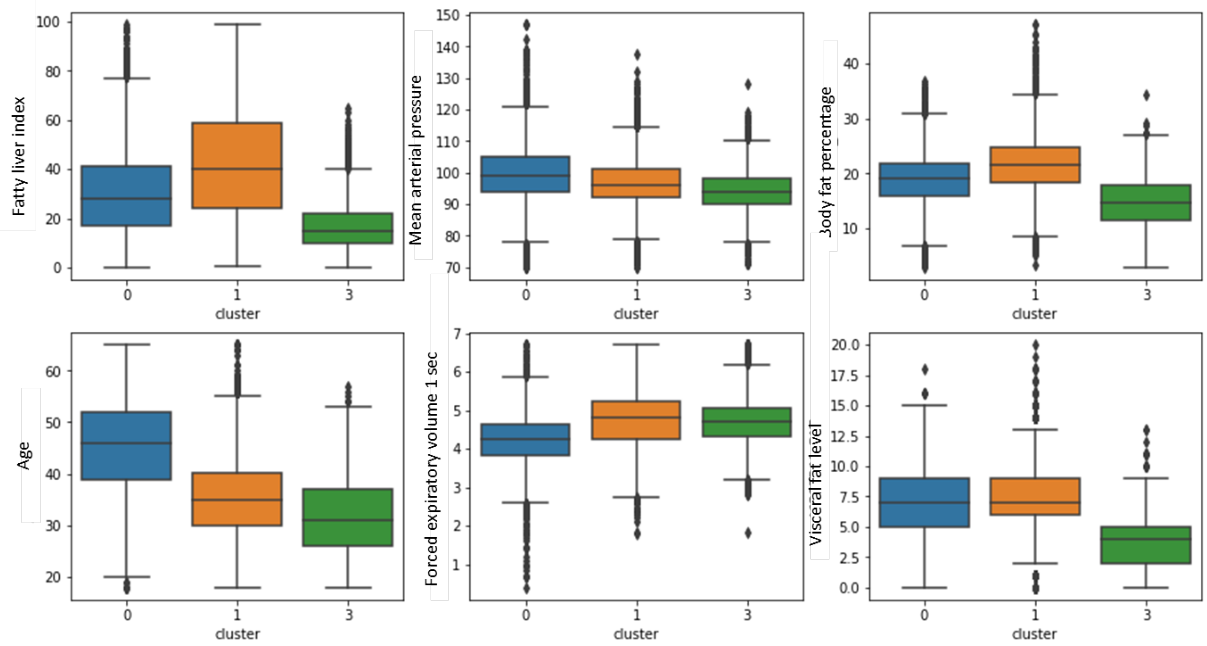
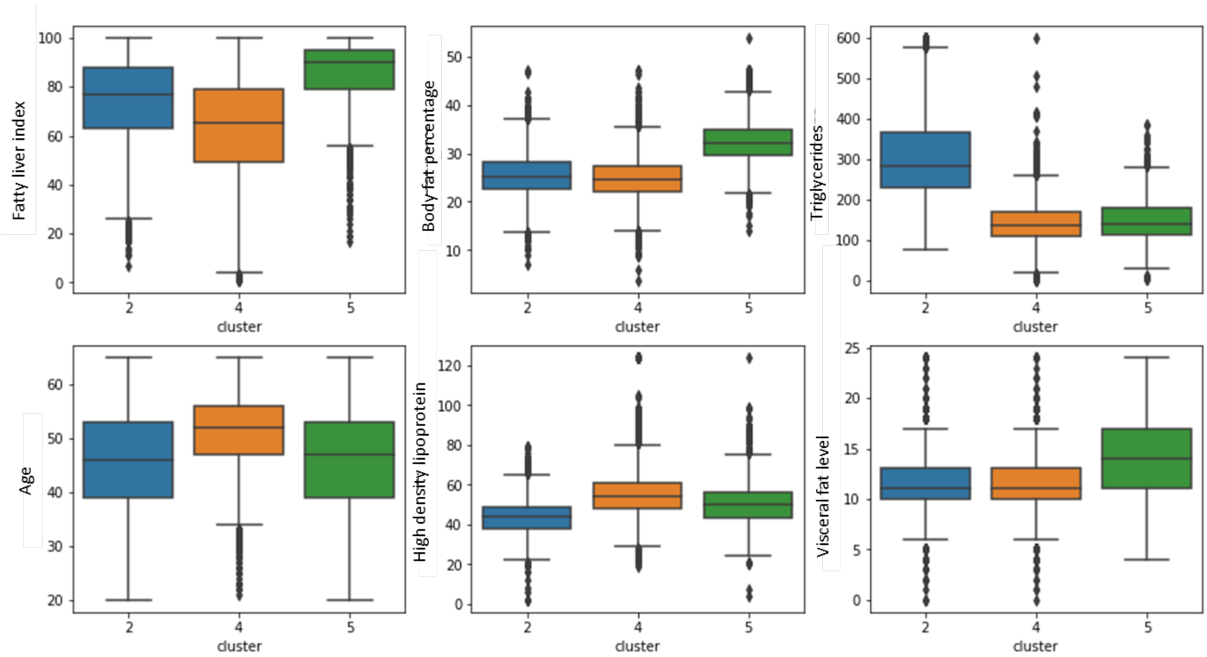


Figure 9: Box plots for high risk groups on males


Fatty liver index
Body fat percentage
Triglycerides
Age
High density lipoprotein
Visceral fat level
cluster