Dimensionality reduction in visualization of high-dimensional mixed data

Alicja Ciuńczyk
STUDENT NUMBER: 2030826

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

prof. dr. Eric Postma
prof. dr. Steffen Pauws
dr. Juan Sebastian Olier Jauregui

Industrial supervisors:

dr. Marine Flechet
dr. Nicola Pezzotti
dr. Gert-Jan de Vries

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
May 2020

Contents

List of Figures

Preface

Working on this thesis was an incredible learning experience, but it would not have been possible without the help of people involved in the process.

I would like to thank dr. Elisabeth Huis in 't Veld and dr. Marijn van Wingerden for presenting me with an opportunity to collaborate with Philips on my thesis project. I would also like to thank dr. Drew Hendrickson for managing the administrative part of this collaboration.

My biggest appreciation is for my supervising team. I don't have enough words to describe my gratitude for prof.dr. Eric Postma, prof.dr. Steffen Pauws, dr. Marine Flechet, dr. Gert-Jan de Vries and dr. Nicola Pezzotti. Thanks to them, I not only learned the theory needed for this project but also how to be a better researcher. Thank you for always being available, answering all my questions, and letting me explore things on my own.

I would also like to thank the whole Philips team for making me feel welcome. The atmosphere in the office, and later in Teams, made the whole experience enjoyable, even when the outside world was not.

Dimensionality reduction in visualization of high-dimensional mixed data

Alicja Ciuńczyk

High-dimensional mixed data require a different approach in the context of dimensionality reduction than homogeneous data. That approach can be either using an algorithm specifically designed for both numerical and categorical variables or by using a distance for mixed data. This research explores the second approach, using t-Distributed Stochastic Neighbor Embedding (t-SNE) combined with Gower distance. The goal of the research was to explore if using t-SNE would allow for better clustering than no dimensionality reduction.

Heart failure data were used with the assumption that it has a separation, according to phenotypes. Dimensionality reduction was used in order to allow for visualization of the phenotypes in two dimensions. To quantitatively find the separation in t-SNE embeddings, a clustering algorithm, DBSCAN, was used. The results of DBSCAN were evaluated by calculating the Silhouette coefficient.

The results showed that t-SNE and data without dimensionality reduction result in very similar clusterings in terms of the number of clusters and Silhouette coefficient. However, t-SNE allows for visualization of the outputs, which allows for additional quantitative evaluation. The visualization of t-SNE results made possible the exploration of the emerging patterns.

1. Introduction

Visualization of data can help in better understanding the world and phenomena within it. However, there is a problem that undermines this visualization. Humans are used to experiencing the world in three dimensions and can only observe and understand a limited number of variables. With a maximum of three axes, researchers can only present a limited number of variables. Many datasets, including clinical data, contain tens or hundreds of variables. Such data, called high dimensional, poses problems for researchers from many fields. Usually, dimensionality reduction can solve those problems by preserving the most important information from the data in two or three dimensions. However, when the high-dimensional data contain both numerical and categorical variables, exploration, visualization, and analysis become harder. This is because the conventional ways of handling high dimensionality problems, like linear Principal Component Analysis (PCA), do not work well when nominal and ordinal variables are combined with numerical ones (Van Der Heijden and Van Buuren 2016). Another reason is that making the data homogenous by discretizing numerical variables or omitting categorical variables can lead to the loss of information, while one-hot or dummy encoding the categories can result in representations that have no meaning (Zhang et al. 2016).

To solve this problem, researchers created several dimensionality reduction methods that can work with mixed data and help with their visualization. Examples of them are Nonlinear Principal Component Analysis (NLPCA) (Gifi 1990) and an extension

of Multiple Correspondence Analysis (MCA) (Greenacre and Blasius 2006). Those two methods are related to each other, and both use indicator matrices to capture the information in categorical variables. A different approach uses a dimensionality reduction technique, for example, t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van Der Maaten and Hinton 2008), with a distance metric that can work with both numerical and categorical variables simultaneously. Each of the approaches has its advantages, as well as disadvantages. In this research, the focus is on exploring the second approach, namely, using t-SNE to gather meaningful insights from high-dimensional mixed data.

High-dimensional data is complex and usually contains tens, hundreds or thousands of variables, which can be on different measurement level (e.g. mixed data), have different units or scale. When dealing with such data, dimensionality reduction can improve the quality of the clustering(Magdalinos, Doulkeridis, and Vazirgiannis 2011; Song, Yang, Siadat, and Pechenizkiy 2013). This enhancement of clustering is due to the fact that many machine learning algorithms, such as clustering, perform better when applied to lower dimensional data, which has reduced complexity compared to high-dimensional ones.

The t-SNE algorithm is a relatively new dimensionality reduction technique. It is quickly gaining popularity (see Appendix A). It was used in research in many different disciplines, as astronomy (Traven et al. 2017), genetics (Li et al. 2017), and chemistry (Fooladgar and Duwig 2019). It is used in medical research to make data visualization possible. For example, Shen et al. (2017) used a variant of t-SNE to visualize the relationships and overlap between diseases. It has also been used to explore phenotypes of different diseases (Hilvering, Vijverberg, Houben, Schweizer, Lammers, and Koenderman 2014; Costantino, Aegerter, Dougados, Breban, and D'Agostino 2016).

According to the World Health Organization (WHO), cardiovascular diseases are a primary cause of death, being responsible for 31% of all deaths worldwide (World Health Organization 2017). One of the risk factors for cardiovascular disease is heart failure (HF), which in 1997 was described as an "emerging epidemic" (Braunwald 1997, p. 1365) and today around 26 million people affected by it (Savarese and Lund 2017). It is a long-term condition in which the heart is not able to pump enough blood to satisfy the needs of the human body. Various subtypes of the diseases have been identified. This makes HF heterogeneous in prognosis, outcome, and response to treatment based on the distinct phenotypes (Faxén et al. 2017). One of the phenotypic classifications is the division between systolic and diastolic HF. In systolic heart failure, the heart does not contract powerfully enough, resulting in reduced ejection fraction of blood from the heart (HFrEF). In diastolic heart failure, the heart does not relax well enough, resulting in insufficient blood filling but maintaining a preserved ejection fraction of blood from the heart (HFpEF). Finding new subcategories and the differences between patients can help to understand the condition better and offer improved patient care. Personalized treatment plans could be beneficial, especially in the light of HFpEF not responding well to 'one size fits all' treatments (Shah, Katz, and Deo 2014; Ferreira et al. 2017; Bertsimas, Orfanoudaki, and Weiner 2018). An analysis and visualization of the patients' data could help with making a phenotype analysis by differentiating a new phenotype, finding out about phenotype classes and distribution, assigning patients to a specific phenotype class, and understanding the variation in patient characteristics, prognosis, and outcome. This is why exploration of the methods that can work with high-dimensional mixed data could have a use in medical research.

The technique explored in this research, t-SNE, can make it possible to see the patterns hidden in high-dimensional mixed data. Such visualizations might be useful for clinicians interested in heart failure phenotypes. It might also improve the clustering

results due to the reduced data complexity. This resulted in the research questions: “Does t-SNE give rise to better clusterings of mixed high-dimensional data compared to clustering on data without the performed dimensionality reduction? What are the factors that can affect the peformance of t-SNE? Can the use of t-SNE help improve the identification of distinct phenotypes in patients with heart failure by allowing the visualization of the data?".

Based on these questions three hypotheses arise:

1. The previous work on enhancing clustering performance with dimensionality reduction suggests that t-SNE provides better clusterings of mixed high-dimensional data compared to the ones done without the dimensionality reduction.
   - The clustering is considered better when it has a higher Silhouette coefficient.
2. The factors affecting the performance of clustering based on the t-SNE output are the parameters of the algorithms and variables included in t-SNE.
3. The use of t-SNE allows and improves the identification of distinct phenotypes in patients with heart failure.

To answer the research questions and test the hypotheses a dataset of patients with heart failure was used, which was high-dimensional and contained both numerical and categorical variables.

2. Related Work

2.1 Clustering

Dimensionality reduction is not a necessary step in working with-high dimensional data. An example of that is research by Shah et al. (2015). It shows that just clustering can help find new phenotypes. So far, this is the only research that used clinical data to investigate heart failure phenotypes. It was done on patients with HFpEF in order to explore possible subtypes in this phenotype. The researchers used hierarchical agglomerative clustering without any dimensionality reduction.

Clustering is used to separate visual patterns in the data quantitatively. A good description is presented by Estivill-Castro (2002, p. 65): "Given a data set, any clustering [...] is a hypothesis to suggest (or explain) groupings in the data". The goal of clustering is to group similar points into a homogeneous cluster based on a predefined criterion while excluding dissimilar points from that group. It is done in an unsupervised way, focusing on the data themselves and thus without using labels.

There are many clustering algorithms with traditional techniques, including clustering based on model, partition, hierarchy, distribution, and density, as well as graph, fractal, and fuzzy theory (Xu and Tian 2015). They differ based on the assumptions about the data, and all of them have advantages and disadvantages. The chosen clustering algorithm should consider the problem at hand and be based on assumptions made about the inherent structure of the data (Estivill-Castro 2002). Each method can present a different output, and assessing which solution is better can be subjective (Xu and Wunsch 2008).

In this research, Density-based spatial clustering of applications with noise (DBSCAN) (Ester et al. 1996) was used. DBSCAN is a density-based clustering algorithm. This means that it finds clusters based on high-density areas, which are separated from other dense clusters by low-density parts. To do that, it uses two parameters: epsilon (Eps) and the minimum number of samples (MinPts). Epsilon defines the radius of the potential neighborhood. The Eps-neighborhood of the point p is defined as

$$N_{Eps}(p) = \{q \in D | dist(p, q) \leq Eps\} \tag{1}$$

where D denotes the dataset and $dist(p, q)$ is the distance between the points p and q.

The DBSCAN algorithm defines two main types of points: core and noise. To define the point p as a core one, it has to have a point q in its neighborhood, which has at least the specified minimum number of samples defined by MinPts. The clusters are seen as chains of points where point p_1 is in the Eps-neighborhood of point p_2 when the core point condition is fulfilled. When a point does not belong to any cluster, it is defined as noise.

In many scenarios, for example, exploratory research, the number of clusters (k) is not known a priori. DBSCAN does not require the user to prespecify k, which can be an advantage. It can also find clusters of a non-convex and arbitrary shape. DBSCAN can also classify points as noise, which is different from some of the other clustering techniques. Many other algorithms will classify all the points to some clusters. DBSCAN will consider a point a noise if it does not fit into the criteria of a cluster specified by Eps and MinPts. This is helpful when evaluating the results because the noise can be treated as outliers and not considered.

Using only clustering, which was the approach of Shah et al. (2015), is interesting, showing that it is possible to find phenotypes without dimensionality reduction or vi-

sualizing the different subgroups. However, their work suffers from two shortcomings: (1) for the clustering, they relied on numerical data only, and (2) the researchers did not use dimensionality reduction, and thus they were not able to visualize their results.

2.2 Dimensionality reduction

Dimensionality reduction, with an emphasis on feature projection, can be used to visualize phenotypes, as shown by Hilvering et al. (2014) and Costantino et al. (2016). The goal of reducing dimensionality is to find low-dimensional structures in high-dimensional data that would allow to represent them in two or three dimensions faithfully. Many dimensionality reduction techniques exist, which can be divided into linear and nonlinear. Linear feature projection finds the best linear representation of the data that allows it to map them to low-dimensional space. An example of it is Principal Component Analysis (PCA), created over 100 years ago by Pearson (1901). He described it as a reduction of the number of dimensions in the data by fitting a line or a plane that gives the best representation.

Using dimensionality reduction can visualize the separation of the patients into known and well-described phenotypes and help with better defining newly found ones. Research by Costantino et al. (2016) studied the phenotypic presentation of patients with early inflammatory back pain by using Multiple Correspondence Analysis (MCA) (Greenacre and Blasius 2006) combined with clustering. MCA is specifically designed to work with categorical data but it can be extended to mixed data as well. Costantino et al. (2016) were able to visualize a clear separation of patients into two distinct groups. The groups were confirmatory of the most recent patients' classification and provided insight into the variables separating them. Such a visualization would not be possible without dimensionality reduction.

Another research that used dimensionality reduction and clustering was performed in 2014 by Hilvering et al. They used Nonlinear Principal Component Analysis (NLPCA) (Gifi 1990) to explore asthma phenotypes. The researchers first established the phenotype of each patient by using sputum analysis, a method currently used for this purpose. Independently from those results, they used NLPCA to try to separate the patients into subgroups. Their results showed a close match between the results from sputum analysis and dimensionality reduction. The conclusion was that using just a routine set of clinical data can deliver a non-invasive way of establishing each patient's asthma phenotype.

MCA and NLPCA are designed for categorical data and mixed data. However, it is also possible to use a different type of dimensionality reduction with a distance that is designed for mixed data. A family of dimensionality reduction techniques that can use such distance is manifold learning. Manifold learning, a part of the nonlinear dimensionality reduction family, assumes that the data points lie on or near a nonlinear manifold. It tries to map the data in the low-dimensional space in a way that preserves the high-dimensional properties (Zheng and Xue 2009).

One of the manifold learning techniques is t-Distributed Stochastic Neighbor Embedding, created by Van Der Maaten and Hinton (2008). It converts a distance measure between high-dimensional points into a joint probability distribution P. This probability is a means of representing similarity. Points that are similar and thus close to each other in high-dimensional space should be preferred to be nearby each other in lower dimensions. Ideally, the probability of point x_i being a neighbor of point x_j in high-dimensional space should be equal to the probability of their low-dimensional equivalents y_i and y_j being neighbors in the low-dimensional space.

The t-SNE algorithm has the advantage of being able to address the crowding problem by using Student's t-Distribution in the low-dimensional space. The crowding problem is the result of the lack of space in low dimensions, which prevents an accurate representation of the distances between points (Van Der Maaten and Hinton 2008). Farther points exert a force that brings moderately distant points closer together, which obstructs pattern creation. The heavier tails of Student t-distribution allow for a better representation of the distances between the points and allow for patterns to emerge.

The low dimensional pairwise similarity in t-SNE uses the Student's t-distribution with one degree of freedom. This similarity is given by the formula

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}}. \tag{2}$$

The high dimensional pairwise similarity uses Gaussian probability distribution and is symmetric by setting

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}, \tag{3}$$

with p_{ij} equal to

$$p_{ij} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma^2)}{\sum_{k \neq l} \exp(-\|x_k - x_l\|^2 / 2\sigma^2)}. \tag{4}$$

The t-SNE algorithm minimizes the Kullback-Leibler divergence between the joint probability distributions in the high- and low-dimensional space and its cost function is given by:

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{5}$$

In t-SNE, the number of neighbors is set by using perplexity. Its value usually ranges between 5 and 50 and is defined with the formula

$$Perp(P_i) = 2^{H(P_i)}, \tag{6}$$

and $H(P_i)$ is the Shannon entropy of P_i. The values of perplexity influence the emphasis on either local or a global structure. When the perplexity values are lower, t-SNE focuses more on a few surrounding neighbors and preserves this local structure, while higher values of perplexity look at more neighbors and recreate the global structure.

The used technique, t-SNE, as many other dimensionality reduction algorithms, uses distances or similarities to lower the number of dimensions. One of the problems with mixed data is that a metric that works well for numerical variables may not work well for categorical ones (Zhang et al. 2015). This is because integers can represent the categorical data, but that does not mean they behave in the same way. For example, it is common practice to represent males and females with ones and zeros. However, because this is only a representation, the natural ordering of one being greater than

zero does not apply. It is thus essential to choose a distance that will take this into account. For this research, several distances were considered, such as Heterogeneous Euclidean-Overlap Metric (HEOM) (Wilson and Martinez 1997) and Unified Similarity Metric (Jinyin et al. 2017). The chosen distance was Gower (Gower 1971) as it can work with mixed data because it separates the nominal and binary variables from the numerical ones and applies different distances to them. It was extended by Kaufman and Rousseeuw (2009) to also work with ordinal and ratio variables. The distance between i and j is given by the formula

$$d_{ij} = \frac{\sum_{f=1}^{p} \delta_{ijk}}{d}_{ij} \sum_{f=1}^{p} \delta_{ijk}. \tag{7}$$

The distance d_{ij} is defined separately for numerical and categorical variables. For numerical data, the distance used is normalized Manhattan defined as

$$d_{ijk} = \frac{|x_{ik} - x_{jk}|}{R_k}, \tag{8}$$

where R_k is

$$R_k = \max(x_k) - \min(x_k). \tag{9}$$

For categorical variables the Dice's coefficient is used. Its formula is

$$d_{ijk} = \begin{cases} 1 & \text{if } x_{ik} \neq x_{jk} \\ 0 & \text{if } x_{ik} = x_{jk}. \end{cases} \tag{10}$$

2.3 Evaluation techniques

To help with evaluating the clustering results, researchers created clustering validity criteria that can be used for this purpose. They can be divided into external and internal. External criteria rely on a known variable or label that is assumed to separate the data. The clustering results are compared with the label and provide an estimate of the quality of the clustering. However, clustering is an unsupervised technique, which means that it should not rely on the presence of labels. In cases where labels are not available, internal criteria can help with evaluating the results. Internal validity is entirely based on the data themselves. Each one makes a different assumption about what makes a good or useful clustering.

The internal evaluation indices that are often used are Dunn Index (Dunn 1973), Davies-Bouldin (Davies and Bouldin 1979) and Calinski-Harabasz (Caliński and Harabasz 1974) and Silhouette coefficient (Kaufman and Rousseeuw 2009).

The evaluation indices mentioned, as well as many others, try to capture the separation and cohesion of clusters. This means that their underlying assumption is that the best clustering should have clusters that are well separated from each other, as well as clusters that contain similar points but not dissimilar ones.

A popular choice is a Silhouette Index, well described by Kaufman and Rousseeuw (2009), which measures how similar the points in a cluster are to each other and how different they are from points in other clusters. It is calculated by using the distance

$d(i, j)$ between points i and j by using the formula

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, \tag{11}$$

where $a(i)$ is the mean distance between the points within the cluster C_i and is defined as

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j), \tag{12}$$

and $b(i)$ is the minimum distance between the cluster C_i and other clusters, given by

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j). \tag{13}$$

The Silhouette Index ranges from -1 to 1, with higher values suggesting a more optimal solution. The Silhouette coefficient represent the overall score and in this research it is defined as the mean of the Silhouette Indices for all the clusters.

Evaluation of clustering results can be also done by comparing the outputs of two different setups, such as different algorithms or in the case of this research, data with and without dimensionality reduction. Such a comparison can focus on showing the similarity of different solutions. One of the metrics that can evaluate that is the Adjusted Rand Index (Hubert and Arabie 1985), which is the Rand Index (Rand 1971) corrected for chance.

3. Experimental Setup

3.1 Data

The data used in this study is a secondary analysis of the OPERA-HF, which is an observational clinical study of heart failure patients (Cleland 2013). The study was conducted between 2012 and 2016 by Philips in collaboration with hospitals in the United Kingdom (NHS Hull and East Yorkshires). The study aimed to validate the performance of heart failure risk models available in the literature, as well as to develop a new model that would show better performance (Crundall-Goode et al. 2013).

In total, 1277 patients took part in the study. All the participants had to be at least 18 years old and reside in the local area. They also had to be at one of the study's hospitals and be treated with loop diuretics or have a clinical diagnosis of heart failure. All participants were able and willing to provide their informed consent.

Data collection included two categories of information. The first category was demographic information and included patients' sex, age, race, marital status, education, occupation, and income. The second category was medical information, for example, symptoms, comorbidities, medication, blood test results, electrocardiogram (ECG) and echocardiography results, as well as psychological health.

The demographic information from Table 1 shows that participants were mostly elderly and overweight males that already had a diagnosed heart failure and previously have had a heart attack.

Table (1) *The table shows basic demographic information based on values present in the data, with valid n showing the number of non missing values. The summary for the first four variables is the percent of the cases, for the BMI and Age the summary is the median.*

		Valid n	**Summary**
Sex	Male	773	61
	Female	504	39
Martial status	Married	614	48
	Single	495	39
Heart failure	Yes	1228	96
	No	49	4
Heart attack	Yes	1102	86
	No	175	14
BMI		834	28.5
Age		1277	77.0

3.2 Preprocessing

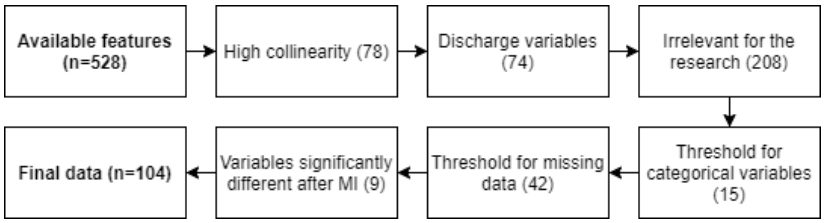
The first step in this secondary analysis, was data preprocessing. It consisted of 3 steps:

- outlier detection,
- feature removal,
- handling missing data.

Firstly, the data were checked for outliers. Values that are not physiologically possible for a human were deemed as errors. The medical literature was used to assess those possible ranges. However, with four variables, a clinician was asked to help with the evaluation. In five cases, the errors were assumed to come from using a different unit, for example, seconds instead of milliseconds. These values were transformed by either multiplying or dividing by a factor of ten. In nine other cases, the values were treated as missing because there was no simple explanation for the value.

The second step was the feature removal, and it was done in phases. Figure 1 summarizes all the phases taken, which will be explained in detail.

Figure (1) *The steps taken in feature removal. The first box shows the number of features available and the last the final number of features. The rest shows how many were removed, the exact number is in the parentheses.*



Firstly, 78 columns were removed due to their high collinearity. This was assessed on three levels: checking for variables that were either an exact match or the opposite of each other, looking at the Pearson correlation coefficients, and using expert knowledge. The variables in question were either a categorical flag of a continuous variable (e.g., a binary variable created from the temperature that indicated if a patient had a fever), expressed a similar occurrence (e.g., hepatitis was rare in the data, which made variables "hepatitis" and "history of hepatitis" identical) or were a result of one-hot encoding the categorical variables in the primary study (e.g., turning variable Sex into two variables male and female).

Only admission features were used to prevent the influence of hospital care on the patients' condition and thus on the phenotype clustering, except if the information was only available at discharge. This meant removing 74 discharge variables.

Next, variables assumed not to influence a heart failure phenotype were taken out of the dataset. This includes information about patients' discharge place, whether their discharge plan was shared with them and their family, the type of hospital at which they were treated, and some demographic variables such as employment or marital status. In total, there were 208 of such features.

Another part of feature removal was ensuring that all the variables had enough categories for train, validation, and test set split. It meant removing variables that contained only one or not enough values to cover the three sets. A threshold was calculated to maximize the probability of each set having at least one instance of each case (the formula for the threshold can be found in Appendix B). The threshold was determined to be a minimum of 13 cases per category. In total, 15 variables were removed due to not meeting the threshold.

3.3 Missing data

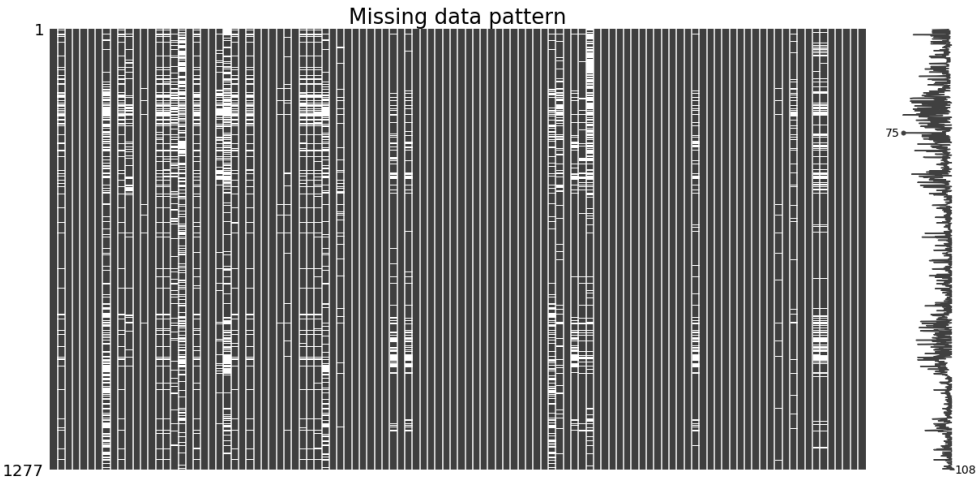
The next phase was focused on handling missingness. Multiple Imputation (MI) technique was chosen as a way of dealing with missing values. It is a highly effective technique, introducing little bias if the assumptions that it operates under are plausible (Pigott 2001). MI is often used combined with statistical modeling, where multiple imputed datasets are used to find optimal model coefficients independently from each other. In the end, their results are pooled together to obtain a final result (Van Buuren 2018). However, in the case of clustering, there is no model for a pooling phase, and it cannot be performed. One solution to this is to stack all the imputed data as proposed by Beesley and Taylor (2019). It creates a large set (number of rows $\times$ number of imputations), which then can be used to perform cluster analysis. Van Buuren (2018) states that this approach can be beneficial when analyzing categorical data. The source further states that it is not a recommended approach because it can produce biased standard errors and t-tests. However, in the case of this research, this concern is not relevant, as the focus is not on estimating population parameters.

The guidelines found in Jakobsen et al. (2017) were used to assess if MI can be performed. The guidelines specify that the proportion of missing data should be between 5 and 40% and that the assumptions of data missing completely not at random (MCAR), as well as missing not at random (MNAR), should not be plausible. In this research, the allowed proportion of missing data was fixed to 40%, which meant that an additional 42 variables had to be removed because they were above that threshold.

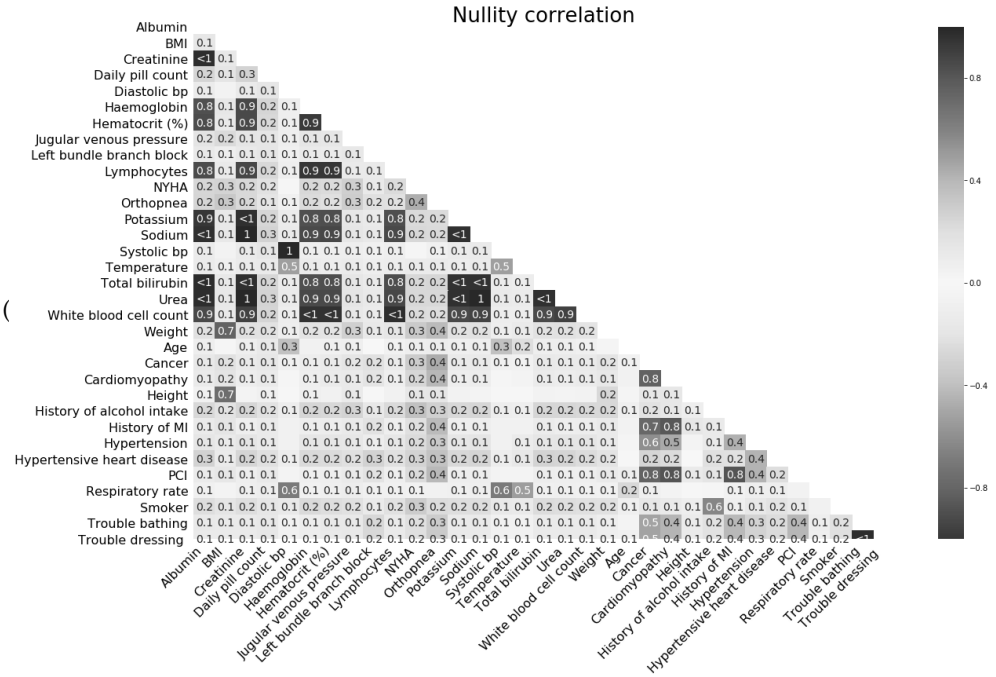
Before performing the MI, the data were randomly split into training, validation, and test sets with the proportion of 60:20:20. In the dataset used for MI, 75 columns had no missing data. The rest of the variables had 15% missing data on average, ranging from 1 to 39%. The missing data patterns and nullity correlation plots (Figure 2) were analyzed in order to assess if performing MI was a reasonable choice.

Figure (2) *The plots used to assess the missing data patterns.*

(a) *Missing data pattern plot. Each row represents a patient in the data. The black parts of the plot indicate present data, while white parts are the missing instances.*



(b) *Nullity correlation plot. It represents the correlation of the missing data between a pair of variables.*



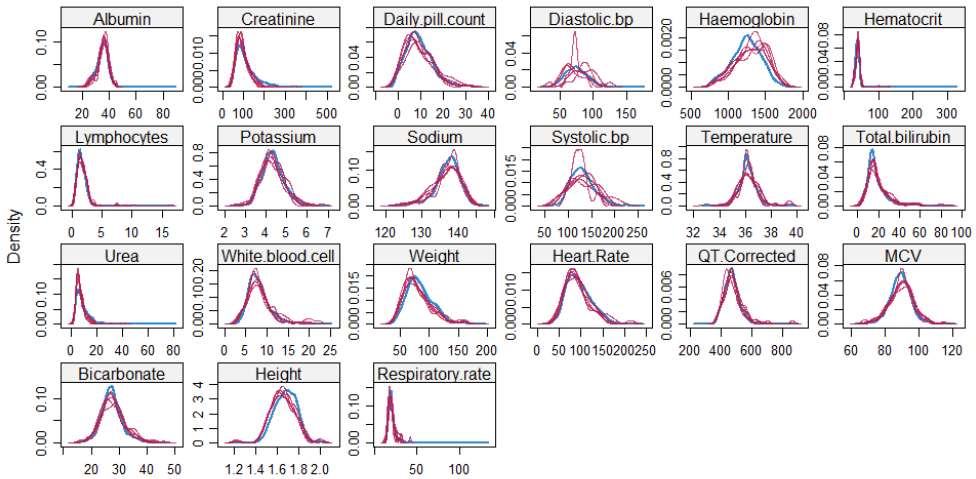
The first plot shows that, in most cases, there are no clear patterns. However, some variables do show them. To better understand where those patterns are originating from, the nullity correlation was plotted. It shows the correlation between the two variables' missing data. Some of the features show a relatively high correlation (above 0.7). Those variables are diastolic and systolic blood pressure taken in one measurement, as well as variables measured with the same blood test (e.g., potassium, sodium, urea). After analyzing the plots, the assumption of MAR seemed plausible and suspected to be related to the test being performed or not.

The Multivariate Imputation by Chained Equations (mice) package in R was used with the default arguments (van Buuren and Groothuis-Oudshoorn 2011). This means that five imputations were performed, as suggested by Van Buuren (2018). The datasets produced by MI were stacked together as described by Beesley and Taylor (2019). This resulted in 4590 rows in the training set and 1536 in both validation and test sets. As for the imputation models, those were also left as default, which means that predictive mean matching (pmm) was used for the numeric data, logistic regression for binary, polytomous regression for unordered categorical and proportional odds model for ordered categorical data. The distribution of the data was evaluated to compare the means and standard deviations of the original and imputed datasets. First, it was done visually by exploring notched boxplots and density plots. For the observed and imputed datasets, the mean, standard deviation, and five-number summary of all the variables were compared to check for differences.

The density plot (Figure 4), in combination with boxplots, showed that two numerical variables needed a closer examination: systolic, and diastolic blood pressure. The curves for the imputations should closely match the observed data. However, those variables showed a visual difference in distribution. The differences in observed and

imputed data were further explored by performing a Welch's t-test, which showed that the differences in means were not significant.

Figure (4) *Density plot for the training data for each of the numerical variables. The blue line represents the observed data, while red lines represent imputations.*



Additional nine features showed a significant difference in means that were assessed using Welch's t-test for numerical and chi-square for categorical variables. They were: Left bundle branch block (p-value 0.009), Arrhythmia (p-value 3.5e-08), PCI (p-value 0.002), Trouble bathing (p-value 1.6e-09), Trouble dressing (p-value 9.0-e10), History of MI (p-value 0.03), Orthopnea (p-value 0.001), Coronary artery bypass graft (p-value 7.8e-05) and History of alcohol intake (p-value 0.03). After closer examination, the conclusion was that the MI did not converge for those variables. However, because they were not used later in the study and had no impact on the final result, they were removed from the data.

This created a final list of variables. It contained 104 features out of which 87 were categorical, with 74 nominal binary, and 2 ordinal variables. The remaining 24 features were numerical, with 18 variables being continuous and 8 discrete. Additional two features were added after performing MI, Body Mass Index (BMI) and Pulse pressure, as they were a combination of features in the data. A full list of variables can be found in Appendix C.

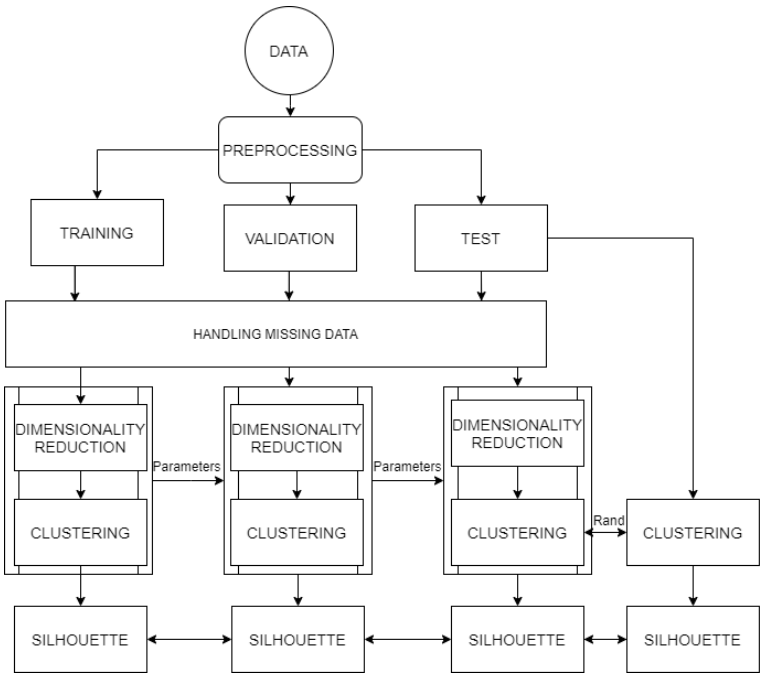
3.4 Method / Models

Figure 5 presents the pipeline used for this research. The first steps, which are preprocessing, splitting the data and handling missing data, were described in previous sections. These steps resulted in the final dataset that was used for the analysis.

As the first step, the subsets of approximately ten variables (further also referred to as features) were created based on medical literature and were used to assess better the feature importance, as well as limit the number of features used by t-SNE.

The training and validation sets were used to perform experiments. The goal of those experiments was to find the highest Silhouette coefficient. There were four sets of experiments performed:

Figure (5) *The pipeline used in the performed analysis. Note: Rand in the figure is referring to the Adjusted Rand Index.*



1. Finding the best performing variable subset;
2. Exploring the parameter influence on the results;
3. Exploring the patterns found in the embedding by looking at variable distribution;
4. Comparing results of t-SNE with the data without dimensionality reduction.

The first experiment consisted of finding the best combination of variable subset and parameters of t-SNE and DBSCAN, which were assessed by the Silhouette coefficient. The algorithms were used on the validation data with the parameters obtained from training. This was done to assess the performance of the solution on unseen data. Secondly, the best performing subset was used to explore and visualize the influence of parameters on t-SNE embeddings, clustering, and the Silhouette coefficient. Next, the best performing subset was used to look at the variable distribution in the t-SNE output. This was done by plotting the t-SNE embedding and using each of the variables in the dataset as a label in that plot. Lastly, a test set was used, where clustering was performed on the results of t-SNE and the test set itself. This was done to explore the effect of t-SNE on the results, as it was the only difference between the two solutions. The Silhouette coefficient was calculated for each result of the clustering (training, validation, and test sets) in order to quantify the performance of t-SNE. The reduced and not reduced test set clusterings were compared by using the Adjusted Rand Index. All their plots were also evaluated visually to catch any discrepancies between the visual patterns and the clustering result. In the next parts, each of the taken steps will be described. The

next step was done using training and validation data to continue with the plan of analysis from Figure 1. These sets were used to find an optimal combination of variable subsets, algorithm parameters, and to test the distance used. To find the optimal variable subset, the medical literature was studied, focusing on heart failure phenotypes. The goal was to match as closely as possible the variables identified as separating different phenotypes. The medical publications used were: Shah, Katz, Selvaraj, Burke, Yancy, Gheorghiade, Bonow, Huang, and Deo (2015); Figueroa and Peters (2006); Clark and Dargie (2011). The subsets based on them will be further referred to as A, B, and C. In many cases, the needed variables were not available in the final dataset. An additional subset (D) was created from the combination of features that were often referred to in the medical literature about heart failure. They are the common tests done when a patient is first seen by a medical professional (American Heart Association 2017). Table 2 shows the exact features identified from each paper.

Table (2) *Subsets of variables that were used in the research.*

Paper/book	Variables
A: Shah et al. (2015)	Age, Pulse pressure, Systolic bp, Diastolic bp, BMI, Potassium, Sodium, Urea, White blood cell count
B: Figueroa and Peters (2006)	Age, Weight, Pulse pressure, Jugular venous pressure, Pulmonary crackles, Pulmonary oedema, Dependent oedema, Liver disease, Pleural effusion, Resting sinus tachycardia
C: Clark and Dargie (2011)	Age, Weight, Sex, Diastolic bp, Heart rhythm, Valvular heart disease, Left ventricular hypertrophy, Haemoglobin, Diabetes
D: Basic tests	Age, Sex, Systolic bp, Diastolic bp, Creatinine, Albumin, Bilirubin, Sodium, Urea, Potassium, Hematocrit, Hypertension

In addition to the four feature sets based on the desk research, the combination of t-SNE and clustering was also applied to the full dataset. On each of the subsets, the Gower distance was calculated, and the resulting distance matrix was fed to the t-SNE and clustered using DBSCAN. For both the t-SNE and DBSCAN parameters were optimized by calculating the Silhouette coefficient and searching for its highest value. For each subset, the perplexity, Eps, and MinPts were optimized to find the highest Silhouette for that subset. The values for the parameters of the used algorithms can be found in table 3 The ranges were intentionally set to be broader than suggested in the literature. The reason is that this is exploratory research, and the influence of parameters on the performance was studied.

Table (3) *Values of the tested parameters for each of the algorithms.*

	Values									
Perplexity	10	20	30	40	50	60	70	80	90	100
Eps	0.001	0.003	0.01	0.04	0.1	0.4	1.4	4.5	15	50
MinPts	10	64	119	173	228	282	337	391	446	500

The outputs of the clustering were also visually assessed; the number of clusters from DBSCAN was compared to the visual number of clusters. The patterns in the best performing subset were explored. Each of the variables from the entire dataset was used as a label in a plot of t-SNE embeddings. The separation of the variables was assessed visually.

The final subset and parameter combination, further called the model, were used on validation data to check if the performance was similar to the training result.

The last step was done using the test data with the final model (t-SNE and clustering), as well as clustering done without the dimensionality reduction. This part of the testing was done to determine the performance of the t-SNE itself. For the data without the dimensionality reduction, the Gower distance was also calculated and the same variable subset was used. This was done to avoid the impact of the chosen distance and subset on the result. The next step was using DBSCAN and finding optimal parameters for it. The Silhouette coefficient was also calculated. However, as the data without t-SNE is high-dimensional, it is not possible to plot it. The two resulting clusterings were compared to each other based on their Silhouette coefficient and the Adjusted Rand Index to assess their similarity.

3.5 Language and version

Most of the research was performed in Python 3.7.4. The general libraries used were pandas (pandas development team 2020), scipy (Virtanen et al. 2020) and matplotlib (Hunter 2007). The most important was scikit-learn (Pedregosa et al. 2011), as functions from it were the most important part of this research. This included splitting the data into training, validation and test set, t-SNE, DBSCAN, Silhouette coefficient, and Adjusted Rand Index.

At the time of this research, Gower distance was not included in any of the big Python libraries. The package gower (Yan 2019) was used to calculate Gower distance.

MI was performed in R. This was done because, at the time of the writing, there were no high-quality Python packages that could perform Multiple Imputation by using different models for numerical and categorical variables[1].

1 The package that could separate numerical and categorical data were Autoimpute (Kearney and Barkat 2019). Unfortunately, at the time, it was not yet a viable solution. During the research, problems with the package were discovered, which led to the imputation models not converging. This resulted in values that were outside of a reasonable range for the variables. The issue was fed back to the author of the python package and solutions suggested.

4. Results

In this chapter, the results of the performed analysis will be presented. The goal of the analysis was to answer research questions and test hypotheses. The research questions focused on the comparison of clustering results with and without using t-SNE, a dimesnionality reduction technique. The research questions also explored the usefulness of t-SNE in phenotype identification. To answer the research question four types of experiments were performed. In this chapter the results of these experiments will be presented. These sets of experiments are: (1) finding the best performing variable subset; (2) exploring the parameter influence on the results; (3) exploring the variable distributions in the t-SNE embeddings; (4) comparing results of t-SNE with clustering done without the dimensionality reduction.

4.1 Variable subsets

As described earlier, the first step of this analysis was exploring results obtained from different variable subsets. They were used to find a solution with the best performance as given by the highest Silhouette coefficient.

The results of the analysis performed on the training data based on variable subsets are in Table 4 (plots for each of the results can be found in Appendix D). This analysis was performed with noise deleted from the Silhouette calculation. From the table, it can be seen that one of the subsets was performing better than others, reaching the Silhouette coefficient of 0.607. The results for all subsets were not changing with the treatment of noise. Whether it was considered one cluster or noisy data were removed, the Silhouette coefficient remained stable.

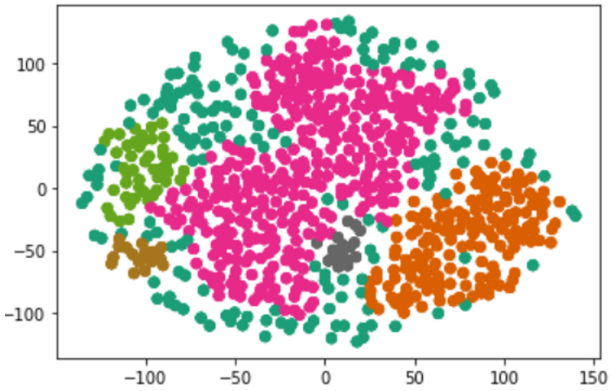
Table (4) *Highest Silhouette coefficients for each variable subset. For each result, the values of perplexity, epsilon (Eps) and the number of minimum points (MinPts) that were used for this score are given.*

Variable subset	Perplexity	Eps	MinPts	Silhouette
A	90	15	173.0	0.370
B	50	4.5	10.0	0.573
C	50	4.5	10.0	0.582
D	40	15	10.0	**0.607**
Full data	10	15	60	0.094

The full training dataset was tested with additional values of epsilon, as after visually inspecting the plot, it seemed that the clustering might not be optimal (Figure 6). With the additional experiments, the full dataset reached a maximum Silhouette coefficient of 0.094. It showed no difference between the results with and without noise.

The best performing combination of a variable subset (D) and parameters from training data identified four clusters (plot d in Appendix D), which is in agreement with the visual assessment. The combination was then applied to the validation data to assess if the solution remained consistent. The validation data showed four clusters and a Silhouette coefficient of 0.516 (the plot for validation data can be found in Appendix E). The distributions of all the variables in the dataset were explored by using them as labels. This showed an interesting pattern emerging, which was further investigated.

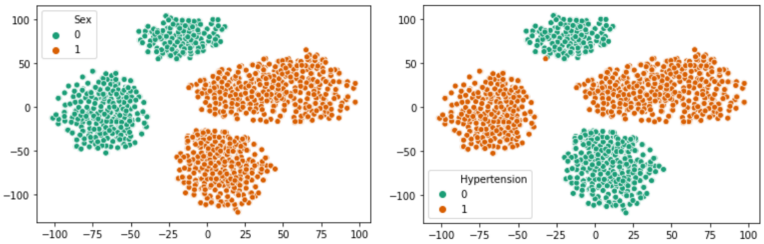
Figure (6) *Clustering result on the t-SNE embedding made form full data.*



4.2 Distribution patterns

To explore the phenotypes within the formed clusters the distribution patterns were explored using training data. Figure 7 shows the subset D plotted with features Sex and Hypertension. This shows an almost perfect separation between the clusters for these two features. It was also confirmed by the chi-square test, where p-values between the clusters with the same category were equal or very close to 1 (e.g., male and male), while the clusters with opposite category had a p-value of 0 (e.g., male and female). As Sex and Hypertension were the only categorical variables in the subset, the question arose if the separation is the result of using categorical features.

Figure (7) *Plots show the perfect separation for the two categorical variables included in the subset.*

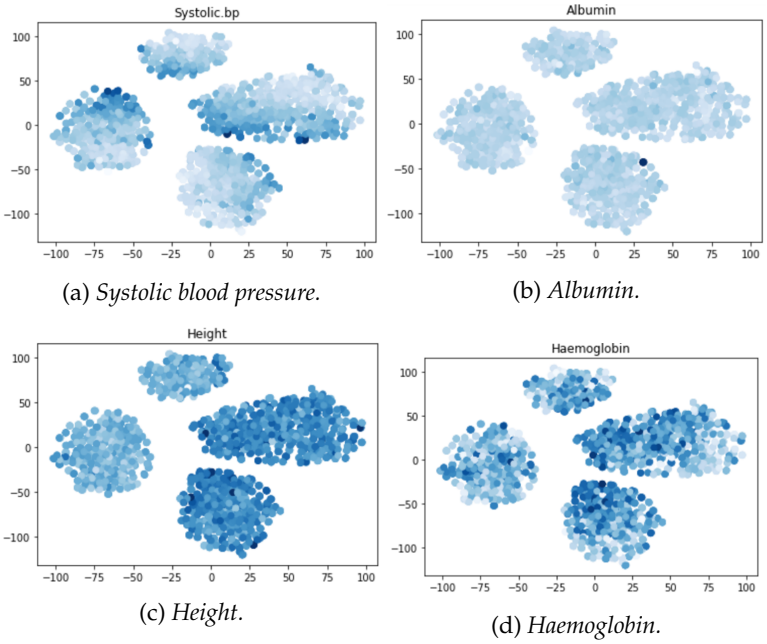


(a) *Plot for the subset D showing the variable Sex.* (b) *Plot for the subset D showing the variable Hypertension.*

Another interesting pattern was seen when using numerical features as labels for the t-SNE plot. What can be seen is a color gradient forming for some of the features within each cluster (Figure 8). For example, when using subset D with four clusters, some of the variables included in the subset will show a gradient within each cluster. It is going from darkest to lightest color and representing the highest and lowest values in the data. There is no clear separation between the clusters in terms of numerical variable distribution like it is seen with categorical features.

To explore this, four experiments were performed using subset D on the training data. As a first step, the variable Osteoarthritis was added to the subset, as it had no visual separation in the previous clustering. It resulted in new clusters forming, which

Figure (8) *Numerical features plotted on the results for subset B. The top two variables were included in the subset, while the bottom two were not.*



(a) *Systolic blood pressure.* (b) *Albumin.*

(c) *Height.* (d) *Haemoglobin.*

showed perfect partition for Osteoarthritis while keeping the separation for the other two categorical features (Sex and Hypertension). COPD, which is another variable that had no separation, was added on top of that. This again created new, ideally separated clusters. In both instances, new clusters were formed, going from the original four to eight and sixteen.

Further, the variables Sex and COPD were removed, which left categorical variables, Hypertension, and Osteoarthritis. This showed four clusters that were separated based on the two features included in the subset, but the effect of Sex and COPD disappeared. The plots for all the experiments can be found in Appendix F.

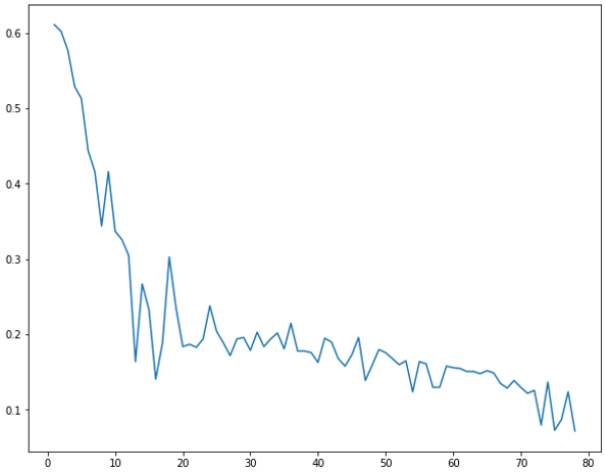
Secondly, the effect of each categorical variable was assessed separately. Each feature was added to the subset, t-SNE, and DBSCAN were applied, Silhouette coefficient calculated, and the feature removed. This meant that each time the subset had two categorical variables (Sex and an added feature). The parameters used were the same as for subset D and were not optimized for each particular combination of variables. This had an effect on the Silhouette, which was not always the highest it could be. However, these experiments were not done to find the best performing subset, but to explore the impact of categorical variables on the t-SNE output and the patterns. Thus, the focus was more on the number of visually separable clusters and the distributions behind them, not on the high performance.

The plots for all of the binary variables showed four clusters forming, usually well separated (examples can be found in Appendix G). They mostly resembled the results obtained from the original subset, and for some of them, the Silhouette coefficient was similar. The only difference between them was the size of the clusters, which could be created by the frequency of classes. However, this would require further investigation, as the sizes of clusters created by t-SNE are not easy to interpret. For categorical, non-

binary variables, the number of clusters forming was related to the number of categories in that feature. For variables with four categories, eight clusters were forming, while those with three categories, ended up with six clusters.

Furthermore, categorical variables were added to the subset incrementally to explore the impact they have on the results. They were added one by one, both in the order they were in the dataset, as well as the reverse until all the categorical variables were included. For this experiment, both t-SNE and DBSCAN parameters were optimized. Figure 9 shows the maximum Silhouette coefficient for each number of variables. With one variable included, the result was higher than the best result for subset D, 0.611 versus 0.607. With the following values added, the Silhouette coefficient began dropping with some spikes in the trend. A similar pattern was seen in the reverse order with several additional spikes.

Figure (9) *Silhouette coefficient value for each perplexity.*



As the last step, the combination of t-SNE and DBSCAN was also optimized for two subsets: one consisting of only numerical and one only of categorical features (Figure H). This was done to see if any clusters would form without the inclusion of any categorical variables. The results of this experiment can be seen in Table 5. The numerical subset, reached a maximum of 0.346 for the Silhouette coefficient, using perplexity of 70, Eps of 15, and MinPts equal to 119. The results were the same, whether the noise was included in the Silhouette calculations or not. It showed one cluster forming with additional noise. The categorical subset had two clusters and noise and reached a maximum Silhouette coefficient of 0.169 with perplexity 80, Eps 15, and MinPts 337. The inclusion of noise did not affect the maximum Silhouette.

Table (5) *Silhouette coefficient and algorithms' parameters for the subset of only numerical and only categorical variables.*

	Perplexity	Eps	MinPts	Silhouette
Numerical	70	15	119	0.346
Categorical	80	15	337	0.169

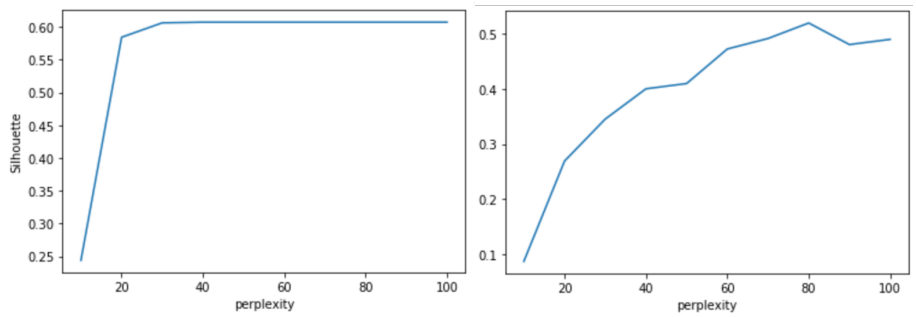
4.3 Parameter effects

The training and validation data were used to find an optimal value for the parameters of t-SNE and the clustering algorithm. It was done for each of the variable subsets. In this section, the impact of parameter optimization is explored.

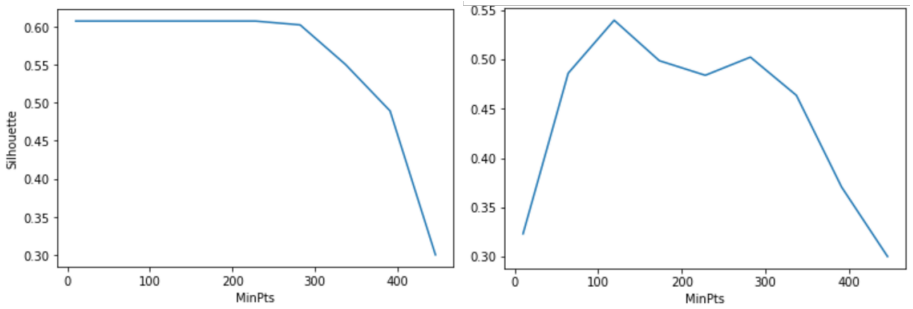
As established in section 4.1, the training data had the best performing combination of parameters on subset D, which showed four clusters, both using DBSCAN and visually. On that subset, 98.8% of cases, epsilon values other than 15, resulted in DBSCAN finding only one cluster in the data. For Eps equal to 15, the average Silhouette coefficient was equal to 0.515. The other 1.2% of the cases had Eps equal to 4.5 and were tied to MinPts of either 10 or 64. Having only one cluster meant that it was only possible to calculate the Silhouette coefficient when the noise was present in the clustering. In the cases where the noise was not present, the values were treated as missing.

Plot *a* in figure 10 shows the performance for all of the tested perplexity values on the training data. For each value of perplexity, the maximum (left plot) and the average

Figure (10) *Silhouette coefficient for the tested parameters. Plots on the left show the maximum performance, while plots on the right show the average Silhouette for that value. Clustering results with only one cluster forming were removed.*



(a) *Silhouette coefficient for each perplexity value.*



(b) *Silhouette coefficient for each MinPts value.*

(right plot) Silhouette coefficients are recorded. The maximum Silhouette was stable for perplexity values between 30 and 100. Knowing that the results of the t-SNE at each perplexity value are visually similar and have the same maximum value and adding to that the epsilon showing more than one cluster almost entirely for the value of 15, the rise of the average needed more exploration. This exploration showed that with higher perplexity values, a larger range of MinPts values resulted in maximum performance.

For each value of MinPts, the same approach as for perplexity was used. The maximum (left plot) and the average (right plot) Silhouette coefficients are shown in plot b in figure 10. The changes in performance for each value of MinPts showed stable results for the values between 10 and 228. For higher values, the Silhouette coefficient was decreasing (left part of plot b in Fig 10). The average performance for MinPts peaked at the value of 119 and then decreased for higher values.

4.4 No dimensionality reduction

In the final step, test data were used both with and without dimensionality reduction to assess the performance of t-SNE. The reduced data were used the same way as in training and validation steps. The Gower distance was used with and without t-SNE, DBSCAN was applied to those results, and the Silhouette coefficient was calculated for both solutions.

For the model with t-SNE, the results were similar to training and validation. The Silhouette coefficient was 0.511, and 4 clusters were identified. The plot was also assessed visually, and it resembled the results from training and validation models (plot b in Appendix E).

For the model without t-SNE, the results were similar to the ones with dimensionality reduction. The highest Silhouette coefficient was 0.5, with Eps and MinPts equal to 1.4 and 10. Four clusters, as well as twenty-four cases of noise, were identified. The results for this model did not differ whether the noise was included in calculating the Silhouette or not. The Adjusted Rand Index value was very high, 0.98, which shows that the results with and without dimensionality reduction were almost identical.

5. Discussion

The research question in this analysis focused on the exploration if t-SNE give rise to better clusterings of mixed high-dimensional data compared to clustering on data without the performed dimensionality reduction and possible factors influencing that. Moreover, the potential of t-SNE in improving the identification of distinct phenotypes in patients with heart failure was explored. Based on these questions three hypotheses were formed:

1. The t-SNE algorithm provides better clusterings than the same procedure without dimensionality reduction.
2. The parameters and variables influence the performance of t-SNE.
3. The use of t-SNE can help with phenotype identification.

To answer the research questions and test the hypotheses four categories of experiments, described in the Results section, were performed: (1) finding the best performing subset of variables; (2) comparing the results of clusterings from t-SNE and not reduced data; (3) exploring the influence of the parameters on the results; (4) investigating the patterns in the t-SNE embedding by looking at variable distribution to determine if they can be distinct phenotypes.

In the first category, the training data were used to find the subset of variables that would give the highest Silhouette coefficient. This subset was the "Basic test" set (D), which included variables: Age, Sex, Systolic, and Diastolic blood pressure, Creatinine, Albumin, Bilirubin, Sodium, Urea, Potassium, Hematocrit, and Hypertension. Two variables (Sex and Hypertension) are categorical. The final performance of this subset on the training data gave a Silhouette coefficient of 0.607 with perplexity 40, epsilon 15, and a minimum number of samples equal to 10. These parameters were applied to the validation data, which showed similar results. Four clusters were identified, and the Silhouette coefficient was 0.516. The same parameters were used on the test data. The results were similar to the validation, four clusters were identified, and the Silhouette coefficient was 0.511. DBSCAN identified four clusters which was consistent with the visual assessment.

In the second set of experiments, the test data were also clustered without t-SNE. The Gower distance was used, and DBSCAN parameters were optimized. The final results showed a Silhouette coefficient of 0.5 and an Adjusted Rand Index equal to 0.98.

The third category of experiments was exploring the influence of parameters on t-SNE by using the Silhouette coefficient. Perplexity between 30 and 100 gave the same maximum result. For epsilon, in most cases, the only value that gave more than one cluster was 15. The number of minimum samples was stable for values between 10 and 228.

The fourth category was focused on analyzing the patterns found during the training phase. The clusters showed perfect separation for any set of categorical variables included. Each result formed clusters to accommodate every possible combination of all the categories. Moreover, numerical variables included in the subset were showing gradients within each cluster. The variables that were not included in the subset did not show such gradients.

5.1 Pipeline results

One of the results of this research was creating a pipeline for the analysis. It can be used to assess any dimensionality reduction technique, provided that the clusters in the data are approximately spherical. It can also be modified by changing the clustering or validity criterion to accommodate non-convex clusters.

By using the pipeline, it was possible to evaluate performance for each of the steps. The best training result showed a Silhouette coefficient of 0.607; for validation set, it was 0.516 and for test set 0.511. Due to the mentioned limitation of this study, it is difficult to assess if it is a relatively good result, but it is possible to say if one solution was better than another. However, looking back at the Silhouette definition, it is possible to assess if the points within the clusters are closer to each other than in other clusters. That holds in this research, as the coefficient is higher than 0.5. On average, the points are more similar within the cluster.

DBSCAN identified four clusters for each of the sets, which was in agreement with the visual assessment. It was also the case with data without the dimensionality reduction. The results for test data clustered without the t-SNE embeddings were similar to those with the reduced dimensions. The Silhouette coefficient differed by 0.011. This small difference shows that the hypothesis that t-SNE provides better clusterings of mixed high-dimensional data compared to the clusterings done without the dimensionality reduction, is disproven by the results of this study.

Despite the similarity of the results, it can be argued that t-SNE, while being an additional step, is an integral part of the pipeline. It allows for visualizing the results, which has an added advantage. This advantage is that visualization can act as an additional evaluation. This is especially important when using clustering as part of the evaluation. In some cases, as with the full dataset, the clustering might not seem visually optimal. Another possibility is that the clustering with the highest validity criterion is not the one that is visually the best. In those cases, the researcher can adjust the parameters or choose a more appropriate evaluation method. That would not be possible without plotting the data.

Another aspect is that visualization helps with understanding the meaning behind the patterns. In the research, no phenotypes were identified. However, the meaning behind the patterns was explored. Plotting the results allowed for exploring an intriguing pattern found in the data. Thanks to that, it was possible to inspect the influence the variables in the subset had on the final result. Using only descriptive statistics and t-test would not directly point to that occurrence.

5.2 Parameters and t-SNE

The second category of experiments investigated the effect of parameters on the performance of t-SNE. This was done for both the t-SNE and DBSCAN parameters.

The experiments showed that the value of epsilon for DBSCAN was crucial. For most cases where the value was other than 15, only one cluster was forming. The other DBSCAN parameter, which specifies the minimum number of samples, also influenced the final performance. High values (above 228) were resulting in the Silhouette coefficient not reaching its known maximum and clustering not matching the visual output.

Perplexity values that resulted in the maximum Silhouette score were between 30 and 100. Over half of those values were outside the recommended range (5 - 50). With the recommended values covering the difference of perplexity of 45, the range that resulted in the highest Silhouette can be considered broad. Moreover, higher perplexity

values resulted in more compact clusters. This meant that there were more points in the radius determined by epsilon and the broader range of the minimum number of samples, resulting in the maximum Silhouette coefficient.

These results show that when evaluating t-SNE by using clustering, it is essential to tune the parameters of that algorithm. A non-optimal parameter setting can result in an undesirable evaluation that does not match the assumed good clustering. Thus, it is imperative to choose the clustering algorithm and validity criteria that match the structures within the data and the assumptions about what is considered a good cluster.

5.3 Distance

The third category of experiments explored patterns that were found during training and investigating if they are a distinct phenotype. The perfect separation for all the categorical variables needed an explanation. Four experiments were performed to test the impact of categorical features.

From those experiments, the assumption was formed that the clusters forming and having a clear separation is the result of used distance and not a phenotype separation. Moreover, the gradients forming in numerical data further point toward the distance, not capturing the information in them. The clusters forming were most likely not representative of the structures within the data but an artifact of the used metric. Therefore, the research question about the use of t-SNE helping improve the identification of distinct phenotypes in patients with heart failure, could not be answered in this study. This is because of the bias introduced by the used distance.

Gower, the used distance, normalizes the numerical features by using min-max scaling. It then applies Manhattan distance to those variables. The normalization of the numerical data brings all the features into a unit norm. This is done to remove the impact that features can have on the distance when they have a broader range of values than other variables. As the normalized values lie between zero and one, they will have a smaller distance than binary variables, which can only be one or zero.

One of the crucial parts of this research is handling mixed-data, especially high-dimensional. The distance used, Gower, is one of the proposed solutions for representing mixed data, as well as using them with machine learning algorithms. However, the results from this research show that it can create artifacts in the t-SNE outputs. This problem is most likely not solved by other distances, such as HEOM, as they also normalize the numerical data. However, this would require additional experiments. An option that could be worth exploring is finding optimal weights for features to remove some of the impact of categorical variables.

5.4 Future work

Based on the findings from this research several future work opportunities arise. The first and most important work should be focused on finding a better suited distance metric that can appropriately scale the categorical variables. An example of that can be an exploration of finding optimal weights for the Gower distance. It is also possible to use dimensionality reduction techniques such as MCA or NLPCA, which are designed explicitly for categorical and mixed data. A comparison of t-SNE, MCA, and NLPCA could give more insight into working with mixed data and help find a better approach.

Another possibilty of further research is answering the question of phenotype identification by using machine learning algorithms on patient data. Due to the bias

introduced by the distance, it was not possible to answer that question in this research. Therefore, more work should be done in order to explore this topic.

5.5 Limitations

There are two limitations of this study that should be discussed, as they had a possible impact on the outcome. As stated in chapter 3.3 Missing data, 42 variables needed to be removed because of the proportion of the missing data. Some of these features are usually used to determine the phenotype of heart failure. An example of that is the ejection fraction variable, which had almost 90% of the missing data. The inclusion of these features could allow for better separation in the data and create a more refined framework for analyzing phenotypes, where the final results could be compared with a label assigned to a patient.

The second limitation involves the evaluation in an unsupervised way. This can be done with the use of an additional step, like clustering or presenting results to an expert in the field. The researcher must rely on data and expert knowledge. This has a drawback that the internal validity criteria for clustering have their limitations—most of them judge the clustering and not t-SNE itself. The output of dimensionality reduction can be optimal, but the final validity criterion is low because of non-optimal clustering parameters. There is no perfect solution for this problem, but visualizing the results and evaluating them qualitatively can help assess the match between the embedding and the clustering.

6. Conclusion

The research goal was to assess to what extent t-SNE results in better clusterings of high-dimensional mixed data compared to the data without the dimensionality reduction and which factors contribute to t-SNE's performance. Potential identification of phenotypes by visualizing the data thanks to the use of dimensionality reduction and clustering was also explored. A pipeline created for this study helped with answering the research question. The final output showed that the clustering results did not differ between t-SNE and data without the reduction. The identification of phenotypes was not possible due to the bias introduced by the used distance.

The research question explored the extent to which t-SNE results in better clusterings than the data without the dimensionality reduction. The t-SNE algorithm gives rise to a proper visualization and a likely better human understanding by applying dimensionality reduction in high-dimensional mixed patient data. The clustering performed on reduced data corresponds with the clustering done on the non-reduced data. Hence, t-SNE does not result in adding or losing cluster information. However, dimensionality reduction and the clustering are biased towards the distance metric used. The current distance, Gower, is a sub-optimal solution favoring categorical over continuous variables, which is a purpose for future research.

The other part of evaluating t-SNE is the chosen setup. It is important to remember that the output of t-SNE was assessed by the Silhouette coefficient calculated for the results of the clustering. Dimensionality reduction and clustering come with hyperparameters that need to be optimized for the best possible clustering and visualization. The pipeline and preprocessing construction enable easy experimentation and replication of this optimization process by a grid search method. The key question is how to assess the best possible clustering and visualization. For that, we have chosen for a 'proxy' measure called the Silhouette coefficient, which reliably trade-offs cohesion within a cluster and separation between clusters.

The factors impacting the results of t-SNE were the variables included, as well as parameters of t-SNE and DBSCAN. The perplexity, t-SNE parameter, had a broad range of values that resulted in the highest maximum Silhouette. The visual outputs showed little difference between the values of perplexity. However, the changing DBSCAN parameters affected the final Silhouette. Epsilon, one of DBSCAN's parameters, had the most impact on the results. It resulted in clusters forming for a small range of values. This shows that in order to evaluate t-SNE by using a clustering algorithm, the parameters of that algorithm need to be well-tuned. It also helps that with t-SNE, the results can be visualized and evaluated qualitatively.

An interesting pattern was found when potential phenotype separation was explored. The pattern in the data was determined to be the result of the distance used and was an artifact of categorical data having more impact on the final output. This effect will be present whenever Gower distance is used with mixed data. The t-SNE algorithm definitely adds to a better visualisation of the clustering of mixed patient data which can help in the discovery of phenotype. Experiments with different subsets of patient variables have demonstrated the practical use of t-SNE. However, the bias due to the use of the distance metric requires further study before the work towards phenotyping in heart failure.

Although t-SNE does not give rise to better clusterings than not performing a dimensionality reduction, it can be a helpful tool in research. gives rise to a proper visualisation and a likely better human expert understanding by applying dimensionality reduction in high-dimensional mixed patient data. The clustering in the dimensionality

reduced data corresponds with the clustering of the original data set. Hence, t-SNE does result in adding or losing cluster information in the data. However, the dimensionality reduction and the clustering are biased towards the distance metric used. The current distance, Gower, is a sub-optimal solution favouring categorical variables for continuous variables which is definitely purpose for future research.

References

American Heart Association. 2017. Common tests for heart failure.

Beesley, Lauren and Jeremy MG Taylor. 2019. A stacked approach for chained equations multiple imputation incorporating the substantive model. *arXiv preprint arXiv:1910.04625*.

Bertsimas, Dimitris, Agni Orfanoudaki, and Rory B Weiner. 2018. Personalized treatment for coronary artery disease: A machine learning approach. *Circulation*, 138(Suppl_1):A11213–A11213.

Braunwald, Eugene. 1997. Cardiovascular medicine at the turn of the millennium: triumphs, concerns, and opportunities. *New England Journal of Medicine*, 337(19):1360–1369.

Caliński, Tadeusz and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Clark, Andrew L and Henry Dargie. 2011. *Oxford textbook of heart failure*. Oxford University Press.

Cleland, John. 2013. *Observational study to predict readmission for heart failure patients*. ISRCTN. https://doi.org/10.1186/ISRCTN96643197.

Costantino, Félicie, Philippe Aegerter, Maxime Dougados, Maxime Breban, and Maria-Antonietta D'Agostino. 2016. Two phenotypes are identified by cluster analysis in early inflammatory back pain suggestive of spondyloarthritis: results from the desir cohort. *Arthritis & Rheumatology*, 68(7):1660–1668.

Crundall-Goode, A, KM Goode, A Shoaib, G Geleijnse, JJG De Vries, E Robson, K Dobbs, K Wong, AL Clark, and JG Cleland. 2013. Opera-hf study design (risk arm): an observational study to assess and predict the in-patient course, risk of re-admission and mortality for patients hospitalised for or with heart failure. *Age (years)*, 42(74):65–80.

Davies, David L and Donald W Bouldin. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.

Dunn, Joseph C. 1973. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Estivill-Castro, Vladimir. 2002. Why so many clustering algorithms: a position paper. *ACM SIGKDD explorations newsletter*, 4(1):65–75.

Faxén, Ulrika Ljung, Camilla Hage, Lina Benson, Stanislava Zabarovskaja, Anna Andreasson, Erwan Donal, Jean-Claude Daubert, Cecilia Linde, Kerstin Brismar, and Lars H Lund. 2017. Hfpef and hfref display different phenotypes as assessed by igf-1 and igfbp-1. *Journal of cardiac failure*, 23(4):293–303.

Ferreira, Joao Pedro, Robert J Mentz, Anne Pizard, Bertram Pitt, and Faiez Zannad. 2017. Tailoring mineralocorticoid receptor antagonist therapy in heart failure patients: are we moving towards a personalized approach? *European journal of heart failure*, 19(8):974–986.

Figueroa, Michael S and Jay I Peters. 2006. Congestive heart failure: diagnosis, pathophysiology, therapy, and implications for respiratory care. *Respiratory care*, 51(4):403–412.

Fooladgar, E and C Duwig. 2019. Identification of combustion trajectories using t-distributed stochastic neighbor embedding (t-sne). In *Direct and Large-Eddy Simulation XI*. Springer, pages 245–251.

Gifi, Albert. 1990. *Nonlinear multivariate analysis*. Wiley.

Gower, John C. 1971. A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871.

Greenacre, Michael and Jorg Blasius. 2006. *Multiple correspondence analysis and related methods*. Chapman and Hall/CRC.

Hilvering, Bart, Susanne Vijverberg, Leo Houben, Rene Schweizer, Jan-Willem Lammers, and Leo Koenderman. 2014. The identification of asthma phenotypes by categorical pca: combinatorial analysis of clinical parameters and dysfunctional blood eosinophils. *European Respiratory Journal*, 44(Suppl 58):P3006.

Hubert, Lawrence and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2(1):193–218.

Hunter, J. D. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Jakobsen, Janus Christian, Christian Gluud, Jørn Wetterslev, and Per Winkel. 2017. When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. *BMC medical research methodology*, 17(1):162.

Jinyin, Chen, He Huihao, Chen Jungan, Yu Shanqing, and Shi Zhaoxia. 2017. Fast density clustering algorithm for numerical data and categorical data. *Mathematical Problems in Engineering*, 2017.

Kaufman, L. and P.J. Rousseeuw. 2009. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Statistics. Wiley.

Kearney, Joseph and Shahid Barkat. 2019. *Autoimpute*. Python package version 0.11.6.

Li, Wentian, Jane E Cerise, Yaning Yang, and Henry Han. 2017. Application of t-sne to human genetic data. *Journal of bioinformatics and computational biology*, 15(04):1750017.

Magdalinos, Panagis, Christos Doulkeridis, and Michalis Vazirgiannis. 2011. Enhancing clustering quality through landmark-based dimensionality reduction. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(2):1–44.

Pearson, Karl. 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.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pigott, Therese D. 2001. A review of methods for missing data. *Educational research and evaluation*, 7(4):353–383.

Rand, William M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Savarese, Gianluigi and Lars H Lund. 2017. Global public health burden of heart failure. *Cardiac failure review*, 3(1):7.

Shah, Sanjiv J, Daniel H Katz, and Rahul C Deo. 2014. Phenotypic spectrum of heart failure with preserved ejection fraction. *Heart failure clinics*, 10(3):407–418.

Shah, Sanjiv J, Daniel H Katz, Senthil Selvaraj, Michael A Burke, Clyde W Yancy, Mihai Gheorghiade, Robert O Bonow, Chiang-Ching Huang, and Rahul C Deo. 2015. Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation*, 131(3):269–279.

Shen, Xianjun, Xianchao Zhu, Xingpeng Jiang, Tingting He, and Xiaohua Hu. 2017. Visualization of disease relationships by multiple maps t-sne regularization based on nesterov accelerated gradient. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 604–607, IEEE.

Song, Minseok, H Yang, Seyed Hossein Siadat, and Mykola Pechenizkiy. 2013. A comparative study of dimensionality reduction techniques to enhance trace clustering performances. *Expert Systems with Applications*, 40(9):3722–3737.

pandas development team, The. 2020. pandas-dev/pandas: Pandas.

Traven, Gregor, Gal Matijevič, Tomaz Zwitter, M Žerjal, Janez Kos, Martin Asplund, Joss Bland-Hawthorn, Andrew R Casey, Gayandhi De Silva, K Freeman, et al. 2017. The galah survey: classification and diagnostics with t-sne reduction of spectral information. *The Astrophysical journal supplement series*, 228(2):24.

Van Buuren, Stef. 2018. *Flexible imputation of missing data*. Chapman and Hall/CRC.

van Buuren, Stef and Karin Groothuis-Oudshoorn. 2011. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.

Van Der Heijden, Peter GM and Stef Van Buuren. 2016. Looking back at the gifi system of nonlinear multivariate analysis. *Journal of Statistical Software*, 73(4):1–8.

Van Der Maaten, Laurens and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*.

Wilson, D Randall and Tony R Martinez. 1997. Improved heterogeneous distance functions. *Journal of artificial intelligence research*, 6:1–34.

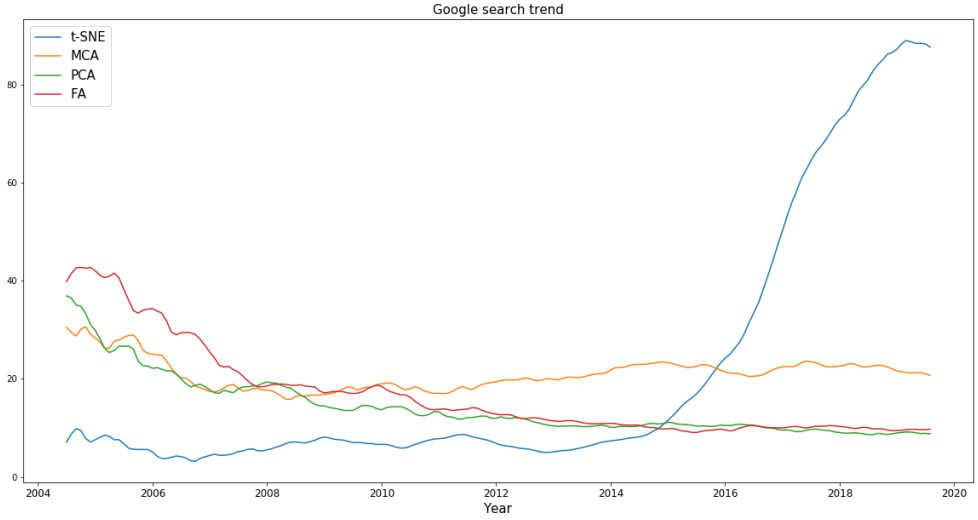
World Health Organization. 2017. Cardiovascular diseases (cvds).

Xu, Dongkuan and Yingjie Tian. 2015. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193.
Xu, R. and D. Wunsch. 2008. *Clustering*. IEEE Press Series on Computational Intelligence. Wiley.
Yan, Michael. 2019. gower.
Zhang, Kai, Qiaojun Wang, Zhengzhang Chen, Ivan Marsic, Vipin Kumar, Guofei Jiang, and Jie Zhang. 2015. From categorical to numerical: Multiple transitive distance learning and embedding. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 46–54, SIAM.
Zhang, Xiao, Changlin Mei, Degang Chen, and Jinhai Li. 2016. Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy. *Pattern Recognition*, 56:1–15.
Zheng, Nanning and Jianru Xue. 2009. Manifold learning. In *Statistical Learning and Pattern Analysis for Image and Video Processing*. Springer, pages 87–119.

Appendices

A. Google search trend

The plot shows trends in Google search results for the whole world. The plot presents t-Distributed Stochastic Neighbor Embedding (t-SNE), exploratory factor analysis (FA), multiple correspondence analysis (MCA), and kernel Principal Component Analysis (PCA). The seasonal decomposition using moving averages was performed on the time series data taken from Google Analytics for the given themes. This plot is the trend part of the results of the decomposition. The data and more information can be found under this link.



B. Proportion of cases calculation

The data were split with the proportion of 60:20:20. This means that there are five buckets with 20% of the data in them. For each categorical variable, the probability was set to 0.95 that at least one from all the categories in those variables will be in training, validation, and test data. The probability of a case not ending up in either of the sets is 0.8. From this point of view, this is a binomial distribution y. Using the formula below, the number of each case in a categorical variable was determined to be higher than 13.

$$\begin{aligned} 1 - (p)^n &= P \\ (p)^n &= P \\ n &= \frac{log(P)}{log(p)} \\ n &= \frac{log(1-0.95)}{log(0.8)} \approx 13 \end{aligned} \tag{B.14}$$

C. Variables

This is a table of all the variables that were available for this research after the preprocessing phase. For each variable, its type is given, either categorical or numerical. The variables in bold are the ones that were used in at least one of the subsets. For some variables, the abbreviation is available in the parenthesis.

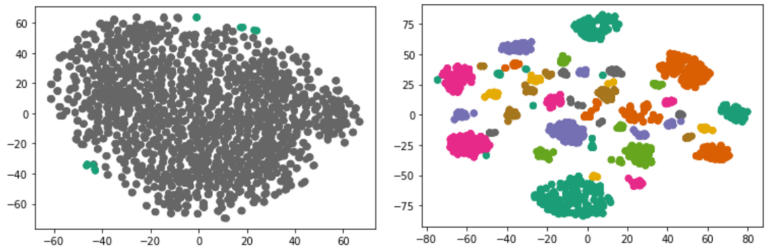
Variable	Variable type
Angiotensin-converting enzyme inhibitor (ACE)	Categorical
Acute coronary syndrome (ACS)	Categorical
Age	Numerical
Albumin	Numerical
Aldosterone antagonist	Categorical
Allopurinol	Categorical
Anemia	Categorical
Angiotensin receptor blocker (ARBs)	Categorical
Antidepressants	Categorical
Aortic aneurysm	Categorical
Aortic stenosis	Categorical
Asthma	Categorical
Body Mass Index (BMI)	Numerical
Beta blocker	Categorical
Bicarbonate	Numerical
Chronic obstructive pulmonary disease (COPD)	Categorical
Cancer	Categorical
Candesartan	Categorical
Cardiac device implant	Categorical
Cardiac arrest	Categorical
Cardiomyopathy	Categorical
Cardiomyopathy or chf	Categorical
Cerebrovascular accident	Categorical
Cerebrovascular disease	Categorical
Clinical signs or symptoms	Numerical
Connective tissue disease	Categorical
Coronary artery disease	Categorical
Coronary heart disease	Categorical
Creatinine	Numerical
Daily pill count	Numerical
Dependent oedema	Categorical
Depression	Categorical
Diabetes	Categorical
Diabetes insulin treated	Categorical
Diabetes insulin treated other	Categorical
Diabetes with end organ damage	Categorical
Diastolic blood pressure (bp)	Numerical
Digitalis	Categorical
End organ damage	Categorical
Gastrointestinal disorders	Categorical
Haemoglobin	Numerical

Heart Rate	Numerical
Heart Rhythm	Categorical
Heart failure (HF)	Categorical
Height	Numerical
Hematocrit	Numerical
History of heart failure (HF)	Categorical
Hypertension	Categorical
Hypertensive heart disease	Categorical
Infection	Categorical
Jugular venous pressure (JVP)	Categorical
Left ventricular hypertrophy	Categorical
Limitation in activity	Categorical
Liver disease	Categorical
Loop diuretic	Categorical
Lung disease	Categorical
Lymphocytes	Numerical
Mean Corpuscular Volume (MCV)	Numerical
Metastatic cancer	Categorical
Myocardial infarction (MI)	Categorical
New York Heart Association heart failure classification (NYHA)	Categorical
Neurological disease	Categorical
Nitrate use	Categorical
Non-metastatic cancer	Categorical
Number of alerting lab values	Numerical
Osteoarthritis	Categorical
Other liver disease	Categorical
Palliative care	Categorical
Peptic ulcer	Categorical
Peripheral vascular disease (PAD)	Categorical
Pleural effusion	Categorical
Pneumonia	Categorical
Potassium	Numerical
Psychiatric disease	Categorical
Pulmonary crackles	Categorical
Pulmonary oedema	Categorical
Pulmonary oedema in notes	Categorical
Pulse pressure	Categorical
QT Corrected	Numerical
Renal disease	Categorical
Renal dysfunction	Categorical
Respiratory rate	Numerical
Resting sinus tachycardia	Categorical
Rheumatoid arthritis	Categorical
ST-T wave changes	Categorical
Sex	Categorical
Smoker	Categorical
Sodium	Numerical
Statin	Categorical
Steroids	Categorical

Stroke	Categorical
Systolic blood pressure (bp)	Numerical
Temperature	Numerical
Thiazide	Categorical
Total bilirubin	Numerical
Transient ischaemic attack (TIA)	Categorical
Ulcer disease	Categorical
Unstable angina	Categorical
Urea	Numerical
Urinary tract disease	Categorical
Valve disease and diabetes	Categorical
Valvular heart disease	Categorical
Weight	Numerical
White blood cell (WBC)	Numerical

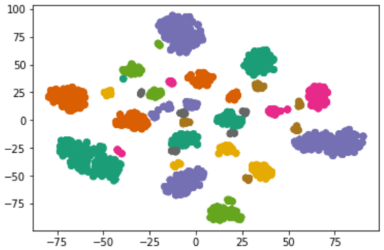
D. Best results for each variable subset

Figure (11) *Best results for each variable subset plotted with the clustering labels.*

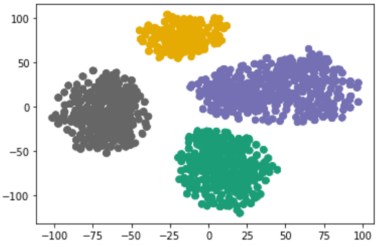


(a) *Subset A.*

(b) *Subset B.*



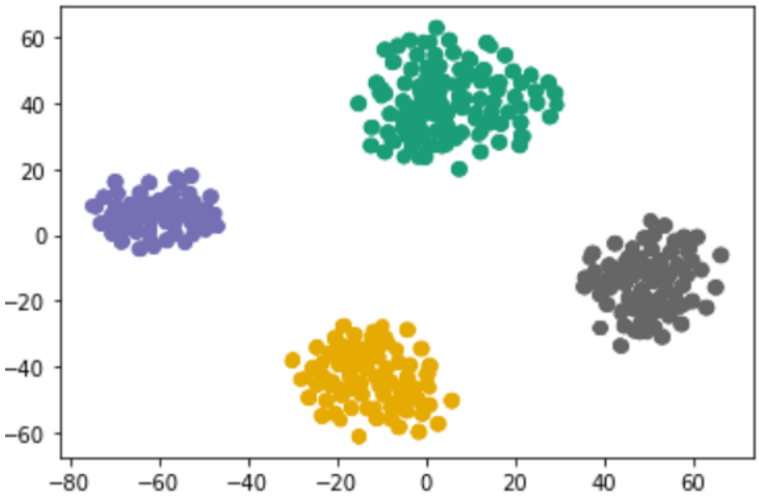
(c) *Subset C.*



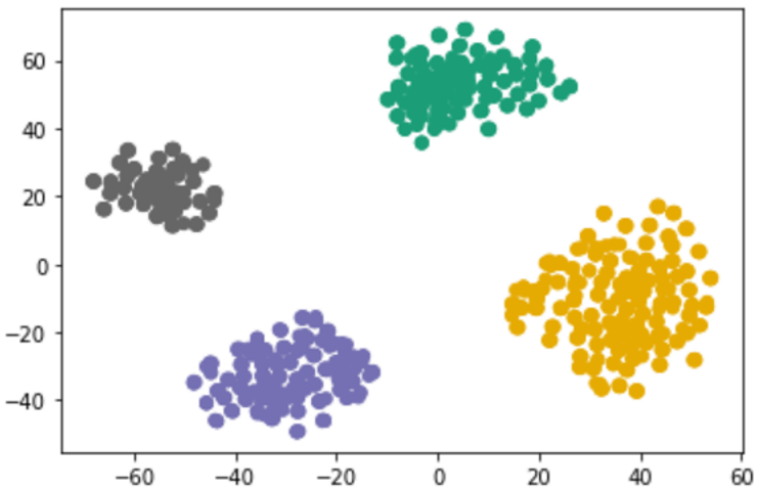
(d) *Subset D.*

E. Validation and test set results

Figure (12) *Results for validation and test sets with each color representing a label from DBSCAN.*



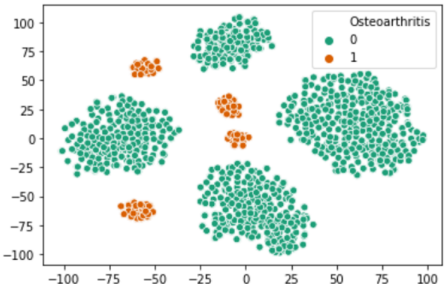
(a) *Validation set.*



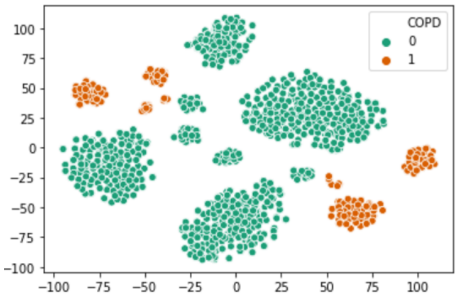
(b) *Test set.*

F. Adding categorical variables

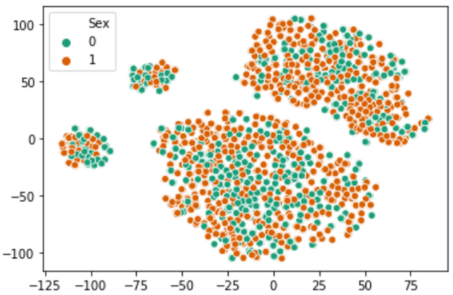
Figure (13) *Plots show the influence that adding a categorical variable has on the output of t-SNE.*



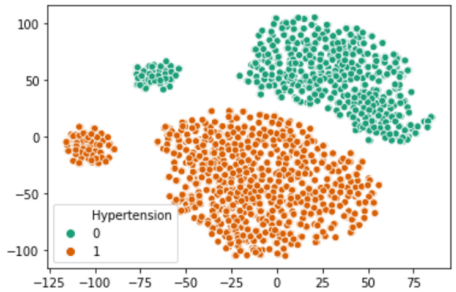
(a) *Plot for the variable Osteoarthritis after adding it to the subset.*



(b) *Plot for the variable COPD after adding it to the subset.*



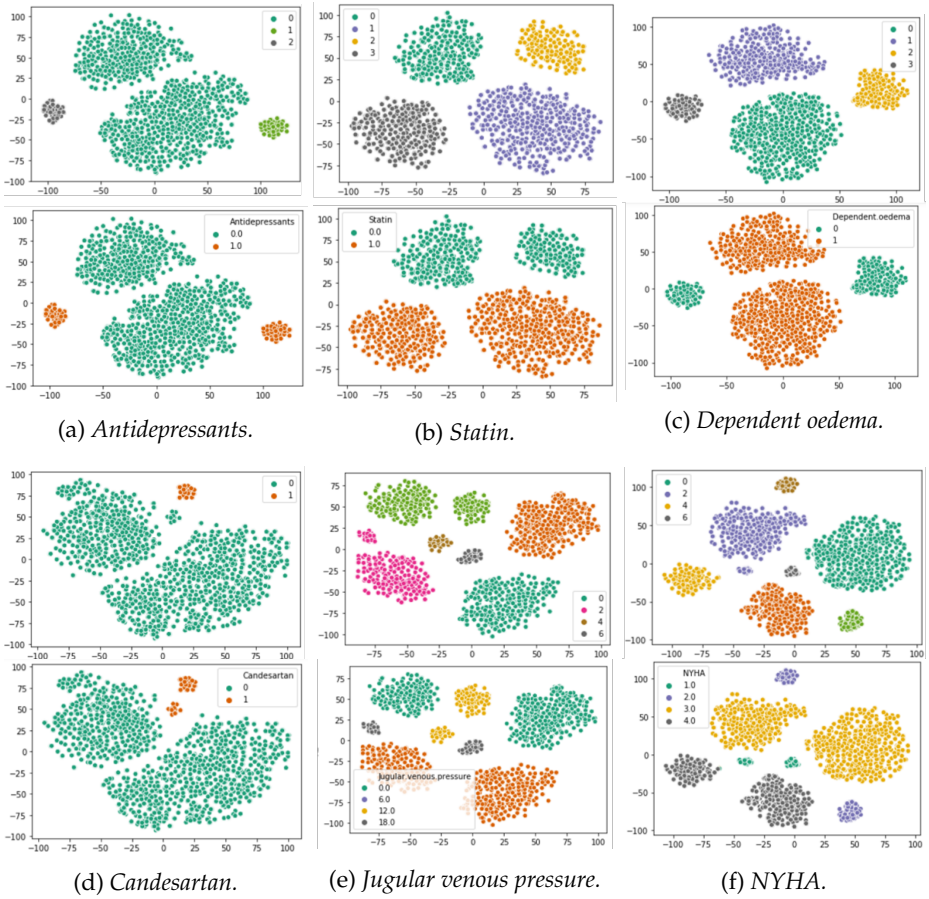
(c) *Plot for the variable Sex with Hypertension and Osteoarthritis in the subset.*



(d) *Plot for the variable Hypertension with Hypertension and Osteoarthritis in the subset.*

G. Variable distributions

Figure (14) *Each of the plots was created by using two categorical variables: Sex and the one depicted on the plot. For each of the plots the top one represents the results of the clustering, while the bottom one shows the distribution of categories.*

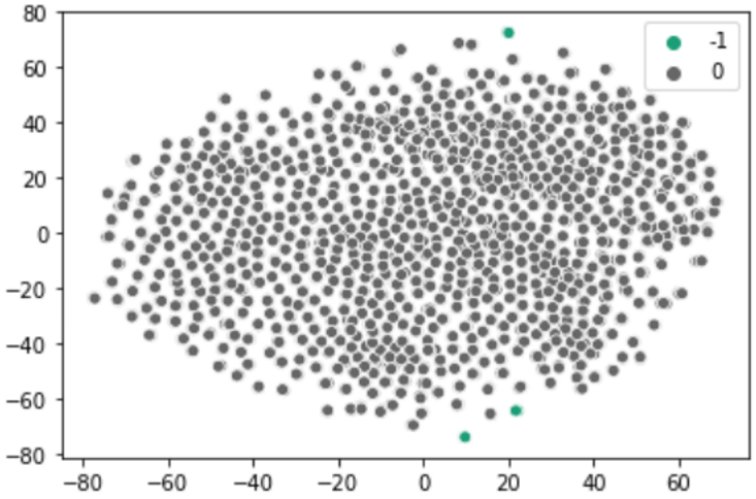


(a) *Antidepressants.* (b) *Statin.* (c) *Dependent oedema.*

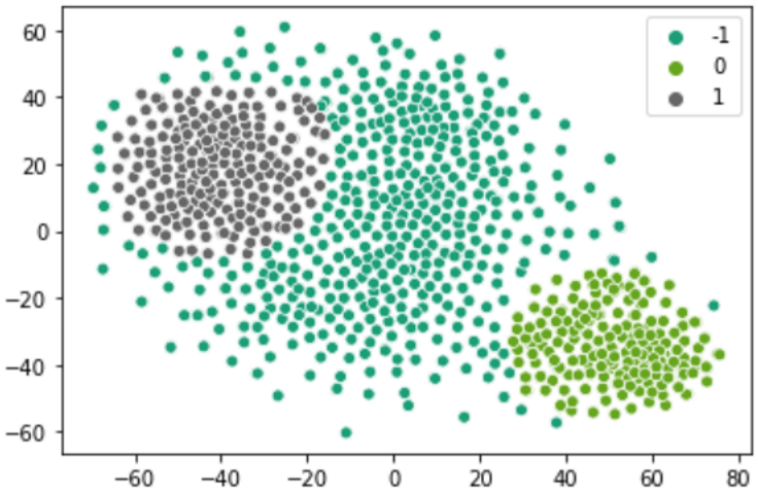
(d) *Candesartan.* (e) *Jugular venous pressure.* (f) *NYHA.*

H. Numerical and categorical variables

Figure (15) *The best results obtained from using only numerical or only categorical variables. Each of the colors represents a cluster defined by DBSCAN.*



(a) *Only numerical variables.*



(b) *Only categorical variables.*