Contents

[1. Research 2](#_Toc44430907)

[2. Plan 3](#_Toc44430908)

[3. Exploratory Data Analysis (EDA) 3](#_Toc44430909)

[Initial Analysis 3](#_Toc44430910)

[Data Cleaning 4](#_Toc44430911)

[Feature Engineering 4](#_Toc44430912)

[Multivariate Analysis 4](#_Toc44430913)

[4. Implementation 5](#_Toc44430914)

[Pre-processing 5](#_Toc44430915)

[Finding the optimum number of clusters 5](#_Toc44430916)

[Create k-means clusters 6](#_Toc44430917)

[Clustering Tendency 7](#_Toc44430918)

[Number of clusters 7](#_Toc44430919)

[Clustering quality 7](#_Toc44430920)

[Generate clustermaps to interpret the clusters 7](#_Toc44430921)

[5. Evaluation 9](#_Toc44430922)

[Objective Evaluation 9](#_Toc44430923)

[Domain Intuition Evaluation 9](#_Toc44430924)

[Patients without Cardiovascular Disease 9](#_Toc44430925)

[Patients with Cardiovascular Disease 10](#_Toc44430926)

[6. Reflection 12](#_Toc44430927)

[Objective 1 12](#_Toc44430928)

[Objective 2 12](#_Toc44430929)

[Discussion 12](#_Toc44430930)

[References 13](#_Toc44430931)

# Research

The chosen dataset from Kaggle is of Cardiovascular Disease patients and can be found at the following page: <https://www.kaggle.com/sulianova/cardiovascular-disease-dataset>.

The dataset is recent since it appears to have been uploaded to Kaggle about a year ago. There are 73 kernels in this dataset with the the top 5 kernels receiving 90/116 votes (77.5%). Since this dataset lends itself very well to Classification, the majority of kernels focus on using various classification algorithms to predict cardiovascular disease (CVD) in patients. There are also some kernels focusing on Exploratory Data Analysis (EDA).

A kernel with 28 votes called “Comparison of Classification – Disease Prediction” is by an author known as Bebab AKCA, a PhD student at University in Istanbul, Turkey. He is currently unranked but does have 4 notebooks and has entered one deep learning competition where he finished in the top 3%. His kernel consists of data cleaning, EDA, feature engineering, model selection, and model evaluation. Focusing on the Machine Learning (ML) aspect, Bebab has split his dataset, normalised the data, and created 5 ML models (decision tree, K-nearest neighbour, naïve Bayes, random forest, and support vector machines). 10-fold cross validation (cv) was used to select the Random Forest model and a grid search (3-fold cv) was used to estimate the optimum hyper parameters. Overall accuracy on the training set was 71.7%. Logistic regression was introduced at this stage and outperformed Random Forests with an overall accuracy of 72.4% on the same training set. Evaluating this model on the test set showed a similar overall accuracy level of 72.7%. The confusion matrix showed a noticeable drop in recall however which is always a trade off with precision. Finally, an Artificial Neural Network was built using TensorFlow which had an overall accuracy of 72.1% on the test set.

In terms of my own interpretation, I downloaded and ran a variation of the kernel without removing any outliers. Bebab assumed that outliers he removed were medically impossible based on his research on google. The overall accuracy on the training set was not significantly different with the exception of naïve Bayes which was reduced by approximately 6% in my variation. My 10-fold cross validation also had almost the exact same results for Random Forest and KNN. I also used a sklearn to generate a neural network model instead of TensorFlow (including 10-fold cv and gridsearch) which also produced an overall accuracy of 73% using default settings.

Another classification kernel called “20 models for Cardiovascular Disease prediction” had 18 votes and its author, Vitalii Mokin is a professor at National Technical University in Ukraine. He seems to be an accomplished Kaggle contributor, being a Competitions expert, datasets expert, notebooks master, and discussion expert. He has authored 49 notebooks and has entered 20 competitions with two bronze awards. He has used the pre-processing from the Bebab AKCA kernel and as the name suggests, he has created 20 ML supervised classification models using the same method above. However, in the early stages, he has run logistic regression to validate assumptions and decisions for feature creating and completing goals. Over the course of running all 20 algorithms, he has used a number of them to create feature importance plots. I noted that XGBoost showed age, and BMI to be the 2 most important factors by far but Light GBM showed ap\_hi, cholesterol, and age to be the top 3 factors. In Vitalii’s kernel, Random Forest was also the most accurate model overall on the test set with a score of 71.56%.

The kernels tend to drop off in terms of votes after this point and some of the contributions are repeated versions of the above kernels. However, a kernel was submitted by an unranked data science student in the US, Elham Amini, on 16/04/2020 and already has one vote. The kernel has some interesting traits such as feature engineering and some interesting questions being asked such as “What are the risk factors associated with cardiovascular disease” which uses statistical tests to answer them. This was an interesting twist on the same problem and could be something interesting to explore further.

The most popular kernel is entirely EDA which is not of much interest from an ML point of view. EDA, data cleaning, and feature engineering tasks are better addressed in the kernels mentioned above.

Overall this dataset is in an interesting topic, has a significant amount of ML classification work and EDA performed, and has some domain knowledge and feature engineering performed which could be used in future ML work. However, classification would not be an area of interest since it has been very dominant in almost all kernels so far.

# Plan

Cardiovascular disease (CVD) is the leading cause of death in Ireland and worldwide (1). Understanding the risk factors that lead to this disease could help to identify groups most at risk. My objectives for this project are as follows:

1. To use clustering to identify which groups of people are most at risk of CVD. These clusters could show different segments of patients most at risk which could be used for targeted interactions with patients for awareness, screening, and treatment of CVD.
2. To use clustering to show the profiles of people that do not have CVD, hence could be used to promote healthier lifestyle choices.

None of the kernels in Kaggle address this aspect of CVD. Some kernels do list important factors but have not expanded this idea to show different profiles of cardiovascular disease patients. Most clustering kernels on Kaggle seem to focus on classic problems such as customer segmentation, finance, and entertainment rather than healthcare.

I plan to use k-means clustering to identify distinct clusters and hierarchical clustering to visualise the results in a clustermap (heatmap). Feature engineering will also play a key role in the project and I plan to use not only some of the work from existing kernels but also test whether other pre-processing methods such as principle component analysis could improve results.

Other factors that need to be considered would include a decision on the number of clusters, size of clusters, and of course interpretation of the results which may require some basic domain knowledge of Cardiovascular Disease.

Building on the research in the other kernels and interpreting the results of this analysis could reveal insights not shown on this dataset up to now. Individual risk factors were identified but we might be able to build on these to show which factors are present at the same time.

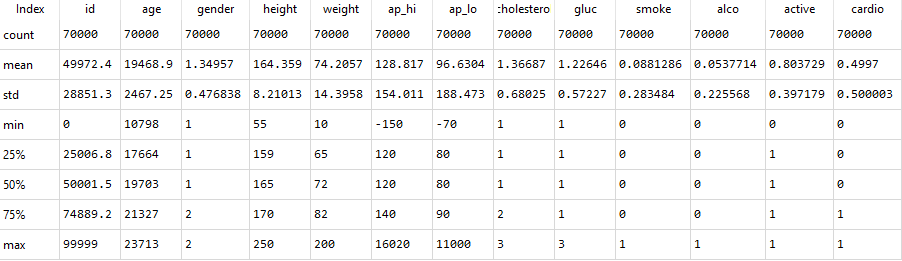
# Exploratory Data Analysis (EDA)

Cardiovascular disease is the leading cause of death in Ireland and worldwide (1; 2; 3). Before clustering is applied to the dataset to generate insights, an understanding of the dataset needs to be gained first.

## Initial Analysis

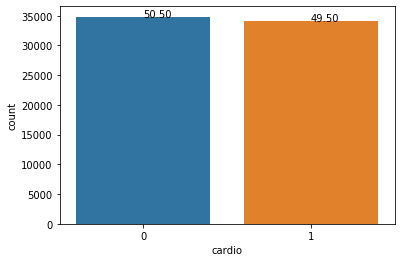
The dataset consists of 70,000 rows and 13 features including target variable which amounts to 910,000 pieces of data. The final binary feature is the target variable and specifies whether the patient has cardiovascular disease or not. The variables are split into three categories:

* Objective variables: factual information such as age, gender, height, and weight.
* Examination: results of medical examination such as blood pressure, cholesterol, and glucose.
* Subjective: information given by the patient such as smoking, alcohol intake, and physical activity.



It was noted that age is the number of days old, height is in cm, weight is in kg, cholesterol and glucose are encoded categorical variables (1 = normal, 2 = above normal, 3 = well above normal). Smoking, alcohol intake, physical activity, and gender (male = 0, female = 1) are all binary variables (4).

Every blood pressure measurement consists of two numbers, one on top of the other. ap\_hi represents the top number and is called systolic blood pressure which is the highest level of pressure your blood reaches when your heart beats. ap\_lo represents the bottom number and is called diastolic blood pressure. This is the lowest level your blood pressure reaches between beats.



Initially from the table above, it looked like the target variable could be imbalanced but from the figure on the left, we can see that the target variable is quite balanced so we will not need to use any over-sampling techniques such as SMOTE - Synthetic Minority Oversampling Technique, for example (5).

## Data Cleaning

All variables are numeric, however, there may be some outliers in the blood pressure variables so negative values were removed as well as values above 250 and 200 for ap\_hi and ap\_lo respectively (6). This totalled 1,039 outliers removed. 24 duplicate values were also dropped from the dataset.

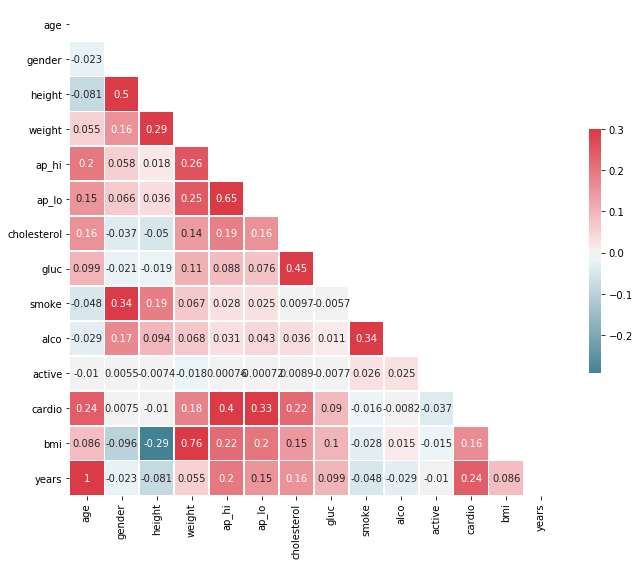
## Feature Engineering

‘Id’ was removed since it does not contribute to the analysis. A number of features were generated such as ‘BMI’ from height and weight and ‘years’ which was the age in days / 365 and rounded off to an integer.

## Multivariate Analysis

To understand which risk-factors have an effect on the target variable, we can use a correlation matrix to show this relationship (4). While the correlation matrix will show correlations between individual features which could be identified as risk-factors, it could be interesting to see how our clustering model groups these risk-factors.

The correlation matrix could also be useful to compare the correlations of our new features such as ‘years’ and ‘bmi’ to the correlations of the features used to create them.

Cardio (the target variable) is correlated with age, weight, bmi, blood pressure (ap\_hi and ap\_lo), cholesterol, and glucose which is expected since these are factual data and examination data generated by clinicians.

Surprisingly, the lifestyle features do not show a correlation with cardio which may point to the subjectivity of this data since it was provided by the patient as a binary value. Perhaps active would yield more correlation if based on data from fitness trackers, for example, or if other questions used were more granular. E.g. if smoking was split into smoker, social smoker, former smoker, never smoked etc.

‘Weight’ is slightly more correlated to ‘cardio’ than ‘BMI’ which may point to weight being slightly better in a predictive model and both ‘years’ and ‘age’ have the same correlation which is not a surprise since ‘years’ was derived directly from ‘age’.

Some basic data cleaning has prepared the dataset for modelling and initial exploration of the dataset has highlighted some initial individual potential risk-factors that can be explored further in our clustering exercise.

# Implementation

The implementation of the clustering exercise was split into 4 stages – pre-processing, finding the optimum number of clusters, building the clustering model, and generating the clustermaps to interpret the results.

### Pre-processing

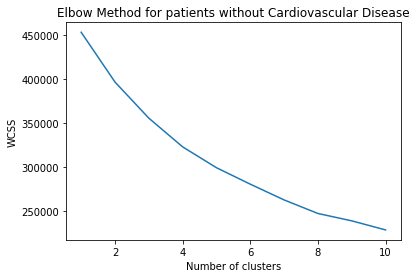
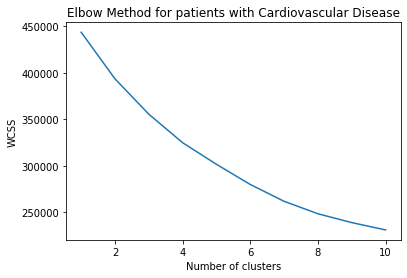
Each instance in this dataset represents a single patient with 13 various features for that patient described as per the previous section on EDA. After EDA, the dataset contained 68,961 instances. To cluster the dataset for patients with and without cardiovascular disease, the dataset was split into two smaller datasets for cardio = 1 and cardio = 0, each containing 34,824 and 34,134 instances. Excluding the target variable, this amounted to 452,751 and 443,742 pieces of data respectively.

To prepare the data for modelling, both datasets were standardised and scaled using StandardScaler from sklearn. When clustering, it is quite useful to plot the results for interpretation. However, when the number of features is greater than 2 or 3, then this becomes a difficult task.

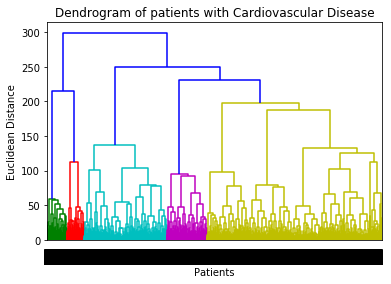
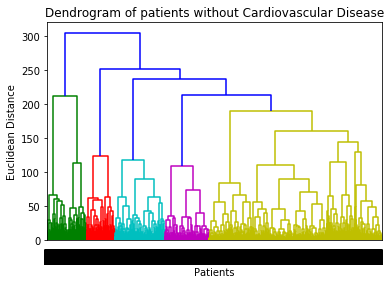
It also may be useful to reduce dimensionality even further using Principle Component Analysis (PCA) to extract the information from the most important variables and help improve the performance of training the model, particularly for hierarchical clustering. PCA was applied to each of the 2 datasets and the results show that in both datasets, 70% of the variance could be explained using the first 5 principle components, and 86% of the variance could be explained by the first 7 principle components. All 100% of the variance was explained in 10 principle components. Since this reduced the number of features in the model, 10 principle components were initially used in the creation of the models to maximise performance.

## Finding the optimum number of clusters

When creating a clustering model, the optimum number of clusters needs to be found. For each dataset, the elbow method was used by creating multiple models with between 1 and 10 clusters where the within-cluster-sum-of-squares (inertia) is calculated for each model. The scree plots that were generated did not have a clear “elbow” point to show the optimum number of clusters.

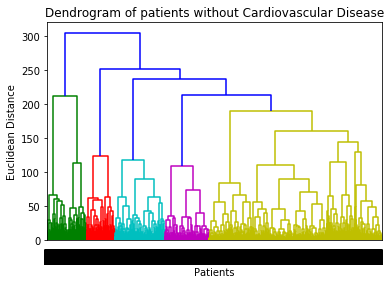
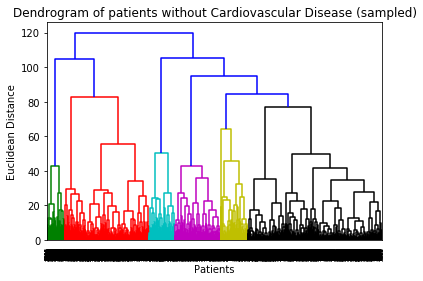


Perhaps around the 7 or 8 cluster point, it starts to level off, but it is not clear. To overcome this issue, a dendrogram was generated for each dataset using hierarchical clustering from the scipy library (method was ward). The results showed that a good number of clusters seemed to be 7 which is around a distance of 150. At 7 clusters, the dendrogram shows the greatest distance between the horizontal lines.



The generation of the dendrograms were very resource intensive in terms of CPU & memory usage and time (over 55 minutes each!), even with dimensions reduced to 10 principle components. To overcome this challenge, a number of different algorithms were attempted including single linkage from scipy as well as calculating distance using pdist from scipy since it performed very well on a benchmark of clustering algorithms (7). Using the fastcluster library (8), the time did reduce to approx. 45 mins.

A random sample of 10%, 20%, 30% of rows were then taken and after a runtime of approximately 6, 12, and 18 minutes respectively, and the results were compared to the dendrogram with the full population:

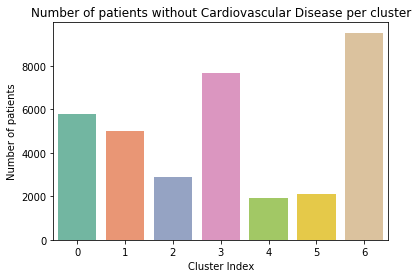
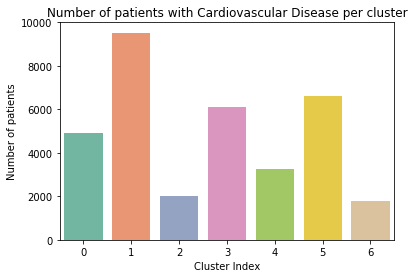


While there were some similarities, the dendrograms from random samples (10%, 20%, 30%) were not accurate enough to use as a method of increasing the performance of the clustering algorithm. Using the fastcluster library was the preferred option since it included all data and reduced the run time to 45 mins (20% time saving). While this was a significant saving relatively speaking, additional efforts would need to be made to further reduce run time using higher spec hardware or distributed computing.

## Create k-means clusters

Once the optimum number of clusters were identified as 7, two k-means clustering models were created for each of the two datasets. K-means++ was used in order to speed up the convergence instead of randomly selecting initial cluster centres. The other parameters used were n\_init = 10 (number of times k-means was run with different centroids) and max\_iter = 300 to prevent the algorithm running for too long.

A plot showing the count of patients in each cluster in both datasets was generated to ensure they were reasonably balanced and to check if the clusters showed any differences between the two datasets. The figures below show these two plots:



Interestingly, we can see some key differences between the 2 groups in clusters with index 1, 5, and 6. These clusters could potentially be used to identify profiles of people more likely to have cardiovascular disease.

To evaluate the clusters from an objective point of view, a number of measures can be used:

### Clustering Tendency

This is ensuring that the data are not uniformly distributed and can be clustered. From the k-means clusters figures above, we can see that the clusters contain different numbers of patients and these numbers also differ when split by the cardiovascular disease feature.

### Number of clusters

Using the elbow method was not conclusive as discussed above but using a dendrogram, while resource intensive, yielded an appropriate result of 7 clusters.

### Clustering quality

There are a number of potential metrics to choose from in the sklearn library. Since ground truth cluster labels are not available, metrics such as adjusted rand index, mutual information-based scores, homogeneity, completeness, V-Measure, and Fowlkes-Mallows scores, were not suitable.

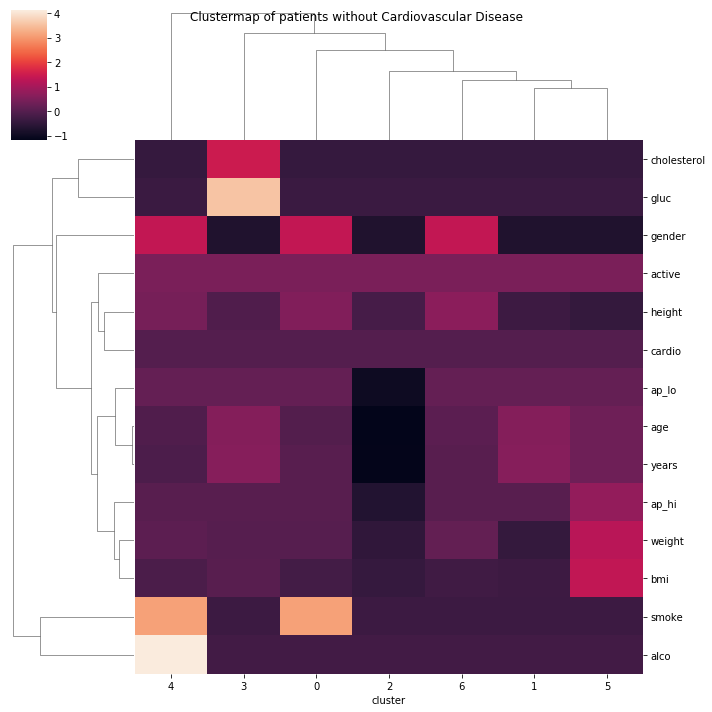
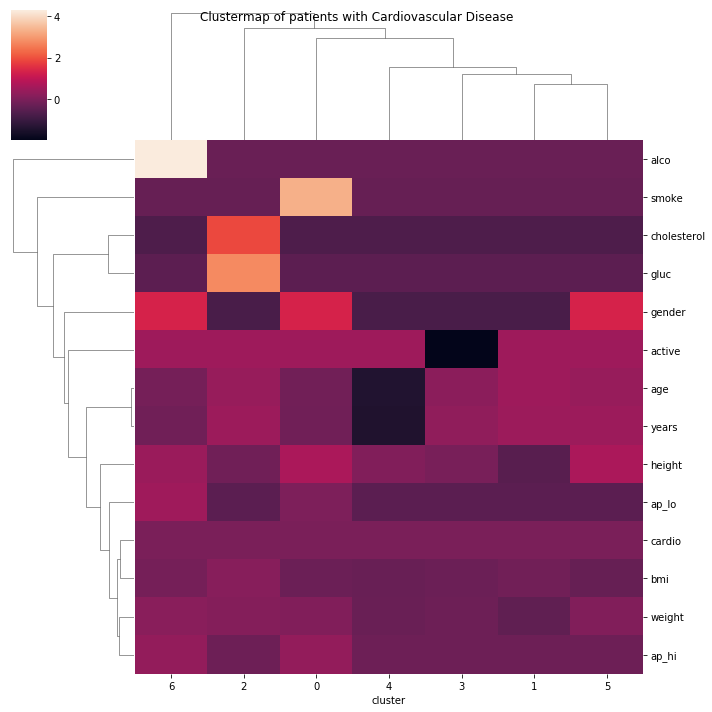
However, the Silhouette Coefficient is appropriate when the ground truth is not known. This has a score of between -1 (incorrect clustering) to +1 (dense clustering). The non-cardiovascular patient dataset and cardiovascular dataset scored 0.17 and 0.14 respectively which indicate that both have overlapping clusters.

An additional metric, Calinski-Harabasz Score, is also appropriate for unknown ground truth labels. A higher score for a model indicates better defined clusters. The non-cardiovascular patient dataset and cardiovascular dataset scored 4,447.2 and 4,021.3 respectively which indicate that the non-cardiovascular dataset has slightly better-defined clusters, but both are close.

Finally, the Davies-Bouldin Index is also appropriate for unknown ground truth labels. A lower score indicates better separation between the clusters (0 is the lowest). The non-cardiovascular patient dataset and cardiovascular dataset scored 1.58 and 1.71 respectively which indicate that both have close scores for separation with non-cardiovascular have slightly better separation.

## Generate clustermaps to interpret the clusters

To interpret these clusters, we can move away from our principle components and see how the clusters vary across the original 13 features of each dataset. Heatmaps are useful for this purpose (9). To prepare the data for the clustermap, a dataframe was created with one row per cluster, containing the mean of each column in that cluster. An additional dataframe was also created containing one row per cluster, with the median of each column, to check if the median would give more distinct separation between variables, in particular the binary variables. The plots generated by median tend to better differentiate between features and are shown in the figures below:



Using the dendrogram of the clusters (columns), we can see that in both cases, 4 groups of clusters emerge. Since the rows are also clustered, the order differs between the two datasets, but this does not affect the interpretation.

Initial interpretation of the clustermaps show that darker colours have a lower value and brighter colours have a larger value. Since the data were standardised, the mean is 0 and the standard deviation is 1 and so all features are on the same scale which means that no feature can dominate the others. We can see that in general, patients with cardiovascular disease tend to have higher values across many of the features compared to patients without cardiovascular disease. In the following section, we will evaluate specific groups of clusters in much more detail to understand the underlying characteristics of these groups of patients.

# Evaluation

## Objective Evaluation

As implemented in the previous section, the evaluation of the clusters from an objective point of view can be measured across the following three factors:

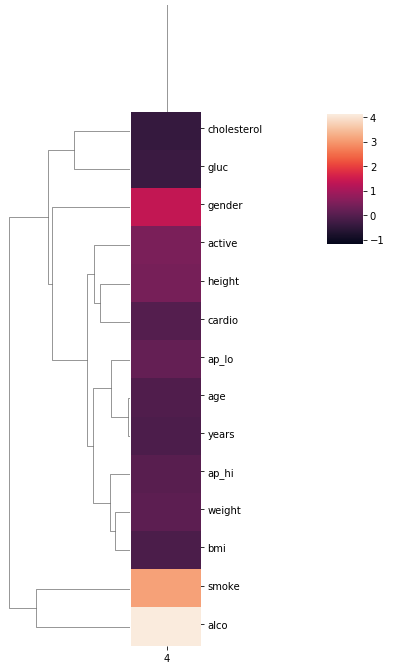
* Clustering Tendency – the clusters figures show that the data are not uniform and can be clustered.
* Number of clusters – using the elbow method was not conclusive as discussed in the previous section but using a dendrogram, while resource intensive, yielded an appropriate result of 7 clusters.
* Clustering quality – Silhouette Coefficient scores of 0.17 and 0.14 show that the clusters are reasonably well separated but are overlapping. Calinski-Harabasz Indexes of 4,447.2 & 4,021.3 and Davies-Bouldin Indexes of 1.58 & 1.71 show that the non-cardiovascular disease dataset has a slightly better separation of clusters, but both are quite close.

## Domain Intuition Evaluation

The cause of cardiovascular disease is not clear but there are many factors associated with increasing the risk of having the disease. The most important of these “risk factors” include high blood pressure, smoking, high cholesterol, diabetes, inactivity, being overweight or obese (BMI > 25), and family history. Other risk factors include age, gender, diet, and excessive alcohol consumption (1; 3; 2). The conclusions drawn from the clustering results should be aligned with the above risk factors and should give an additional level of granularity to specific cohorts of patients affected by or at risk of getting the disease.

### Patients without Cardiovascular Disease

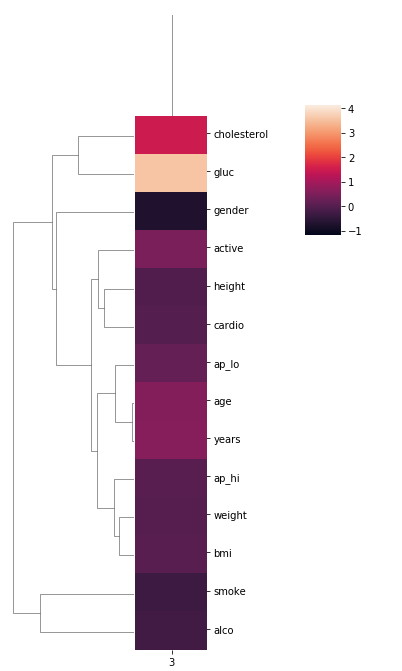
This cohort of patients consisted of 34,827 patients and 4 distinct groups of clusters were identified. They are described as follows:



#### Midlife Females

This cohort of mostly female patients do not have any underlying abnormal results from their medical examinations and report themselves to both drink and smoke. Their BMI seems to be helped by their slightly higher than average height and their age seems to be pretty average (~45 years old).

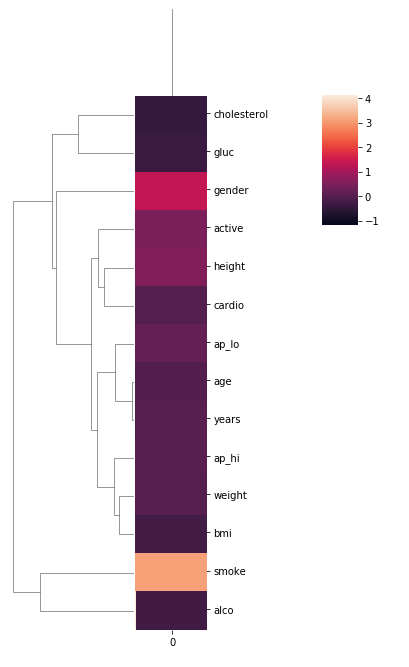
Key messages: patients should quit smoking since it is a major risk factor.



#### Late Midlife Males with warning signs

This cohort of mostly male patients are slightly older than average (>45), do not report that they smoke or drink, and they report that they are active. However, they do have some warning signs that could need to be addressed. This group of patients do have above normal levels of cholesterol and well above normal levels of glucose. While blood pressure does seem to be relatively normal, this could be a warning sign for this group of patients to make some lifestyle changes such as lowering sugar intake or there could be an underlying condition to address.

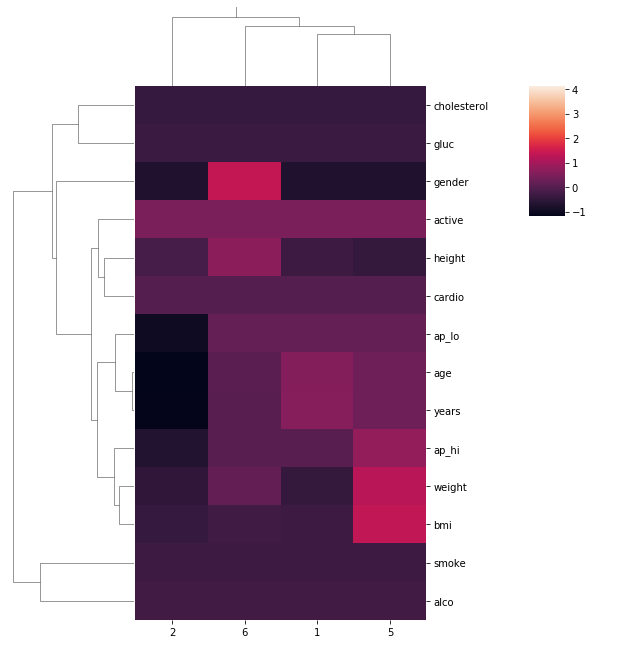
Key messages: patients should lower sugar intake or check for underlying conditions (e.g. diabetes)



#### Female smokers

This group of female patients do report that they smoke but do not drink. They otherwise seem to have pretty normal medical examination readings, they report that they are active, and their BMI seems to be average. Their age seems to be average (~45).

Key messages: patients are generally healthy but should quit smoking.



#### Healthy but early warning signs

This group is mixed in terms of demographics containing about 75% male patients. All report to not smoking or drinking and all report to being active. Cholesterol and glucose levels are normal, however, there is a group of overweight slightly older males with high blood pressure. These males seem to be shorter in height which could account for the higher BMI levels.

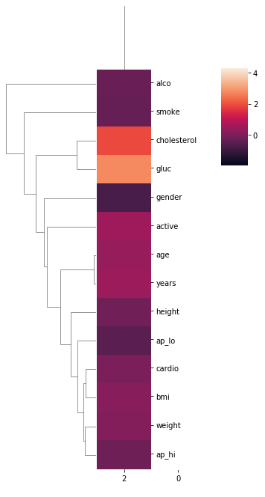
Key messages: patients are generally healthy, but weight should be monitored.

### Patients with Cardiovascular Disease

#### Female Drinkers

This group of female patients are overweight and report drinking but not smoking. They report being active, are slightly older than average (>45) and have higher than average blood pressure (both ap\_hi – systolic blood pressure and ap\_lo – diastolic blood pressure).

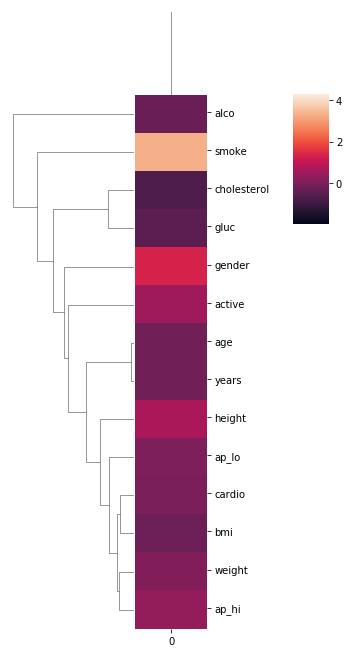
Key messages: patients’ weight should be reduced, and blood pressure monitored.



#### Older males with poor diet

This group of older males’ report being active, not smoking, and not drinking. However, their glucose and cholesterol levels are both well above normal. This could indicate lifestyle issues or an underlying condition. Their diastolic blood pressure (ap\_lo) seems to be close to average but their systolic blood pressure (ap\_hi) seems to be slightly higher than average.

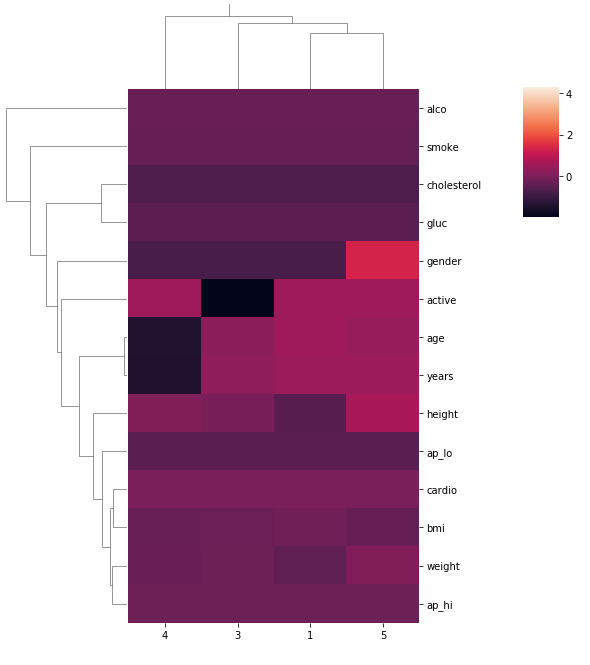
Key messages: patients should reduce weight, reduce sugar intake and check for underlying conditions.



#### Female Smokers

This group of female patients report to smoking but not drinking and being active. Their glucose and cholesterol levels seem to be normal which points to a reasonably healthy diet. Their weight seems to be slightly higher than average and their age seems to be slightly more than average (>45). Smoking seems to be the most dominant feature in this cohort of patients.

Key messages: there is a direct link back to Female Smokers in non-cardiovascular disease patients - this group are at a very high risk of developing the disease.



#### Overweight Patients

There is a mix of patients in this group in terms of demographics, but all share a pattern of being overweight and having higher than average systolic blood pressure (ap\_hi). All report not drinking or smoking, all have normal cholesterol and above normal glucose levels. The group consists of mainly men (~75%) and are mostly older (~75%) patients.

Key messages: across demographics such as age, gender, and lifestyle, it is clear to see that weight and blood pressure are significantly associated with cardiovascular disease in patients.

# Reflection

The two objectives of this exercise were:

1. To use clustering to identify groups of people most at risk of Cardiovascular Disease.
2. To use clustering to show the profiles of people that do not have Cardiovascular Disease, hence could be used to promote healthier lifestyle choices.

## Objective 1

From the clusters shown in the previous section, we can see that key factors such as weight, smoking, glucose, cholesterol, and blood pressure are associated with cardiovascular disease. The clusters group these factors together and it is clear to see the link between lifestyle choices and the medical consequences of those choices. For example, being overweight increases blood pressure, high sugar intake increases glucose and cholesterol levels and leads to obesity, smoking is associated directly with cardiovascular disease in females.

These clusters also show specific sub-groups of patients directly at risk and show patterns of behaviour in specific groups that could be useful in more targeted interactions with patients. For example:

* There is a direct link between women that smoke in their 40s having no cardiovascular disease with women in their 50s that smoke and do have cardiovascular disease.
* Men in their 50s with a poor diet are a sizable cohort of patients at high risk.
* Older men and women that are overweight or obese even if they are reasonably active are very much at risk.
* A sizable number (approx. 8%) of patients with cardiovascular disease are young males that are otherwise healthy, which could point to family history or underlying condition which reinforces the importance of regular screening.

## Objective 2

In general, the levels of cholesterol, glucose, weight, blood pressure, and age were lower in the dataset of patients with no cardiovascular disease. However, some warning signs did appear in this group such as female smokers in their 40s and overweight men in their 50s as discussed above. These warning signs such as smoking and weight also demonstrated the inverse – patients that did not smoke and kept their glucose levels down were not overweight and did not have high blood pressure or high cholesterol. Age is a feature that cannot be changed, and unfortunately older age is more associated with cardiovascular disease. This highlights the need to promote healthier lifestyle in older patients in particular.

## Discussion

Previous work in Kaggle focussed on classification and EDA which was very useful in the early stages of this project to assist with pre-processing, feature extraction, and feature importance. For example, the classification models showed that the most important features were bmi, ap\_hi, age, weight, ap\_lo, and cholesterol (10) and this is aligned with the clustering models in this project. This project built on the work of the classification models and EDA and used clustering to uncover insights that classification would not as easily generate. The purpose of each technique is quite different with classification aiming to predict cardiovascular disease based on the supplied features, and clustering being used to identify homogeneous subgroups in the dataset which can allow awareness, screening and treatment of cardiovascular disease to be more targeted.

In future, this work could be expanded to include additional data which is less subjective (such as patient provided data) or include lifestyle data which is more granular (smoking, alcohol, diet). Additionally, it could be interesting to look at patients from pre-screening through to diagnosis and then following on from early stage to late stage cardiovascular disease to see if lifestyle changes help improve the treatment or outcome.

# References

1. **Health Service Executive.** Cardiovascular Disease. *Health Service Executive.* [Online] 11 07 2011. [Cited: 09 05 2020.] https://www.hse.ie/eng/health/az/c/cardiovascular-disease/.

2. *Obesity and Cardiovascular Disease.* **Lavie, Carl J., Milani, Richard V. and Ventura, Hector O.** 21, New Orleans, Louisiana : s.n., 2009, Journal of the American College of Cardiology, Vol. 53, pp. 1925-1932.

3. **National Health Service.** Cardiovascular Disease. *National Health Service.* [Online] 17 09 2018. [Cited: 09 05 2020.] https://www.nhs.uk/conditions/cardiovascular-disease/.

4. **Ulinova, Svetlana.** EDA cardiovascular data. *Kaggle.* [Online] 2019. https://www.kaggle.com/sulianova/eda-cardiovascular-data.

5. **Srivastava, Vaibhav.** Cardiovascular Disease Complete EDA & Modelling. *Kaggle.* [Online] 04 2020. [Cited: 05 05 2020.] https://www.kaggle.com/vaibhavs2/cardiovascular-disease-complete-eda-modelling.

6. **AKCA, Benan.** Comparison of Classification- Disease Prediction. *Kaggle.* [Online] 01 2020. [Cited: 05 05 2020.] https://www.kaggle.com/benanakca/comparison-of-classification-disease-prediction.

7. **HDBSCAN.** Benchmarking Performance and Scaling of Python Clustering Algorithms. *HDBSCAN.* [Online] 2016. [Cited: 07 05 2020.] https://hdbscan.readthedocs.io/en/latest/performance\_and\_scalability.html.

8. *fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python.* **Muellner, Daniel.** 9, s.l. : Foundation for Open Access Statistics, 29 05 2013, Journal of Statistical Software, Vol. 53, pp. 1-18.

9. **Brunsdon, Chris, Rigby, Jan and Charlton, Martin.** Ireland Census of Population 2011: A classification of Small Areas. *Introduction to R and RStudio.* [Online] [Cited: 08 05 2020.] https://rstudio-pubs-static.s3.amazonaws.com/14998\_c53cef8e65aa4c5386312bc79a0ecb64.html.

10. **Amini, Elham.** CVD: A Complete Analysis (EDA, vis, prediction). *Kaggle.* [Online] 16 04 2020. [Cited: 08 05 2020.] https://www.kaggle.com/eamini/cvd-a-complete-analysis-eda-vis-prediction.