William Edwardo Gunawan 501123334

Heart Failure Data Clustering with Machine Learning

Background

Heart failure is a medical condition in which the heart cannot adequately pump blood, resulting in insufficient oxygen and nutrition reaching the body's tissues and inadequate circulation. For this clustering problem, we will be using the 2020 Heart Failure Clinical Records from UCI Machine Learning Repository.

The dataset comprises clinical records for 299 patients, each associated with 12 clinical features, and it has a data characteristics of multivariate, where it contains both numerical and categorical attributes [1]. Numerical attributes involves quantitative measurements, while qualitative attributes involves qualitative measurements. From the 12 clinical features, we will only use some of the numerical and categorical attributes, which are the following:

> Categorical attributes = sex, diabetes, high blood pressure, smoking, death event
> Numerical attributes = age, serum sodium, platelets

We employ the K-Prototypes clustering algorithm to effectively handle the multivariate data features present in the dataset. This algorithm allows us to cluster different groups of patients based on both numerical and categorical features, providing a nuanced understanding of heart failure subgroups [3].

To determine the optimal number of clusters for meaningful interpretation, we will utilize the Sum of Squared Errors (SSE) curve, applying the elbow method. This statistical approach aids in identifying the point of diminishing returns in cluster improvement, helping us in defining clinically relevant subgroups within the heart failure clinical records dataset.

In summary, I found this dataset fascinating, with the combination of unique multivariate nature and with a reasonable representative sample size of 299 patients, makes it an invaluable dataset for unraveling intricate patterns related to heart failure. The application of the K-Prototypes clustering algorithm, tailored for both numerical and categorical attributes, underscores the sophistication of our analysis. Additionally, by employing the elbow method, we aim to derive clinically meaningful clusters that have the potential to impact patient outcomes against heart failure and contribute to more personalized treatment strategies or improved risk stratification for all patients.

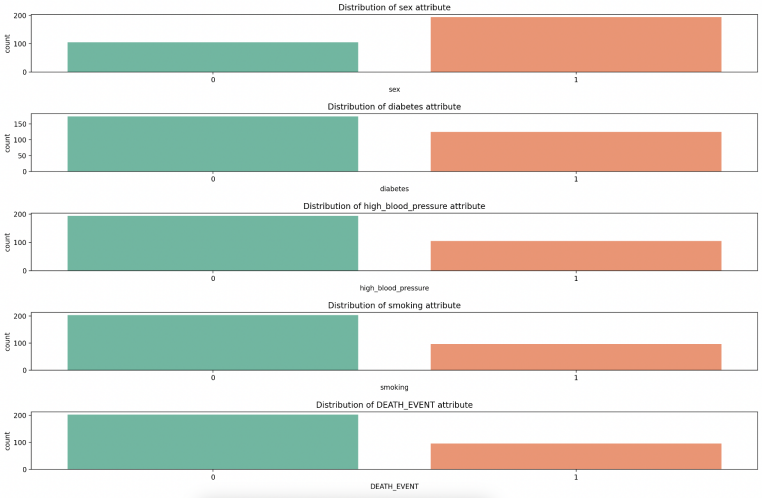
Methods

In this section, we will discuss the methods used for this clustering problem. We will do Data preprocessing, Finding Optimal Number of Clusters by using the Elbow method, and lastly, Clustering the Data.

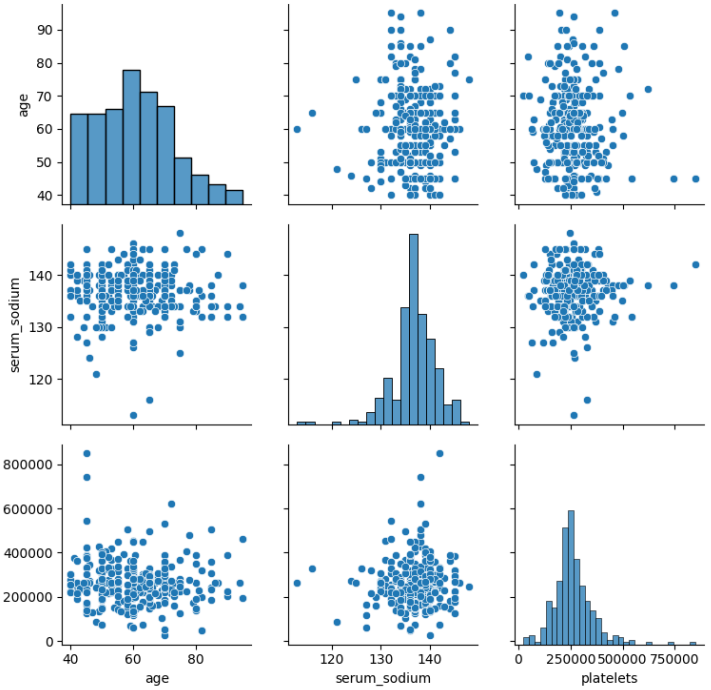
For Data preprocessing, we load the dataset csv file and choose which attributes that we will use. To recall, these are the attributes we chose:

> Categorical attributes = sex, diabetes, high blood pressure, smoking, death event
> Numerical attributes = age, serum sodium, platelets

Categorical attributes before Clustering



Numerical attributes before clustering

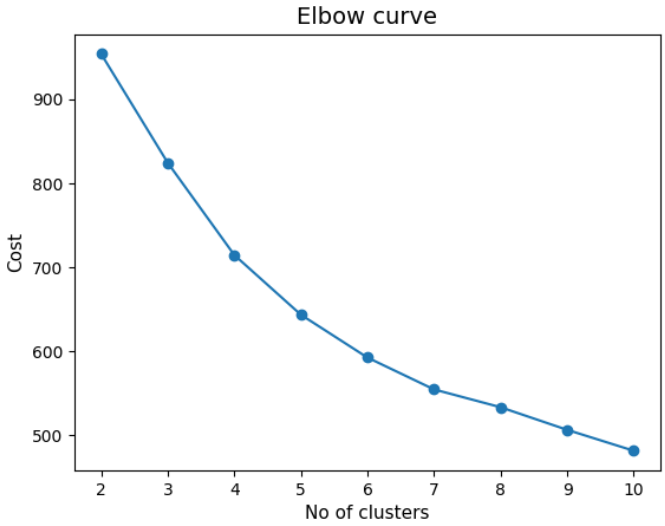


Before performing clustering, we need to scale each of the numerical data features to make sure that each of them is roughly on the same scale, so the clustering is not misleading and biased towards a particular feature and all of features contribute to the clustering process. If we don't scale our feature, features that have larger values will dominate other features. This is the reason why scaling our data is crucial before applying it to a clustering algorithm [2].

After scaling, we will remove the data outliers, since it can introduce noise, affecting the overall structure and cohesion of clusters, and may even lead to the formation of irrelevant clusters. For this clustering problem, we will remove all the outliers from the numerical scaled data that exist outside 3 standard deviations of that features mean. Removing outliers at this stage contributes to a more robust and

reliable clustering outcome. As a result, we were able to remove 6 outliers from the 299 records down to 293 records [5].

Now we will find the optimal number of clusters for our dataset using the SSE curve, the elbow method. The "elbow" in the plot designates a bend that resembles an elbow formed by a sharp change in the rate of decrease of the sum of squared distances. This elbow point, which denotes the point of diminishing returns and beyond which further partitioning of data into clusters does not significantly reduce the sum of squared distances. The Elbow Curve graph suggests that 5 may be the optimal number of clusters for this dataset, as the rate of decrease in the cost function is not substantial beyond this point [4].



Finally, we can start the clustering process using the KPrototypes Clustering algorithm. We will use 5 as the number of clusters for our datasets, and we can see that all 293 records must have a cluster number assigned.

First 5 records

	age	sex	diabetes	high_blood_pressure	serum_sodium	platelets	smoking	DEATH_EVENT	cluster_number
0	75.0	1	0	1	130	265000.00	0	1	4
1	55.0	1	0	0	136	263358.03	0	1	0
2	65.0	1	0	0	129	162000.00	1	1	1
3	50.0	1	0	0	137	210000.00	0	1	0
4	65.0	0	1	0	116	327000.00	0	1	1

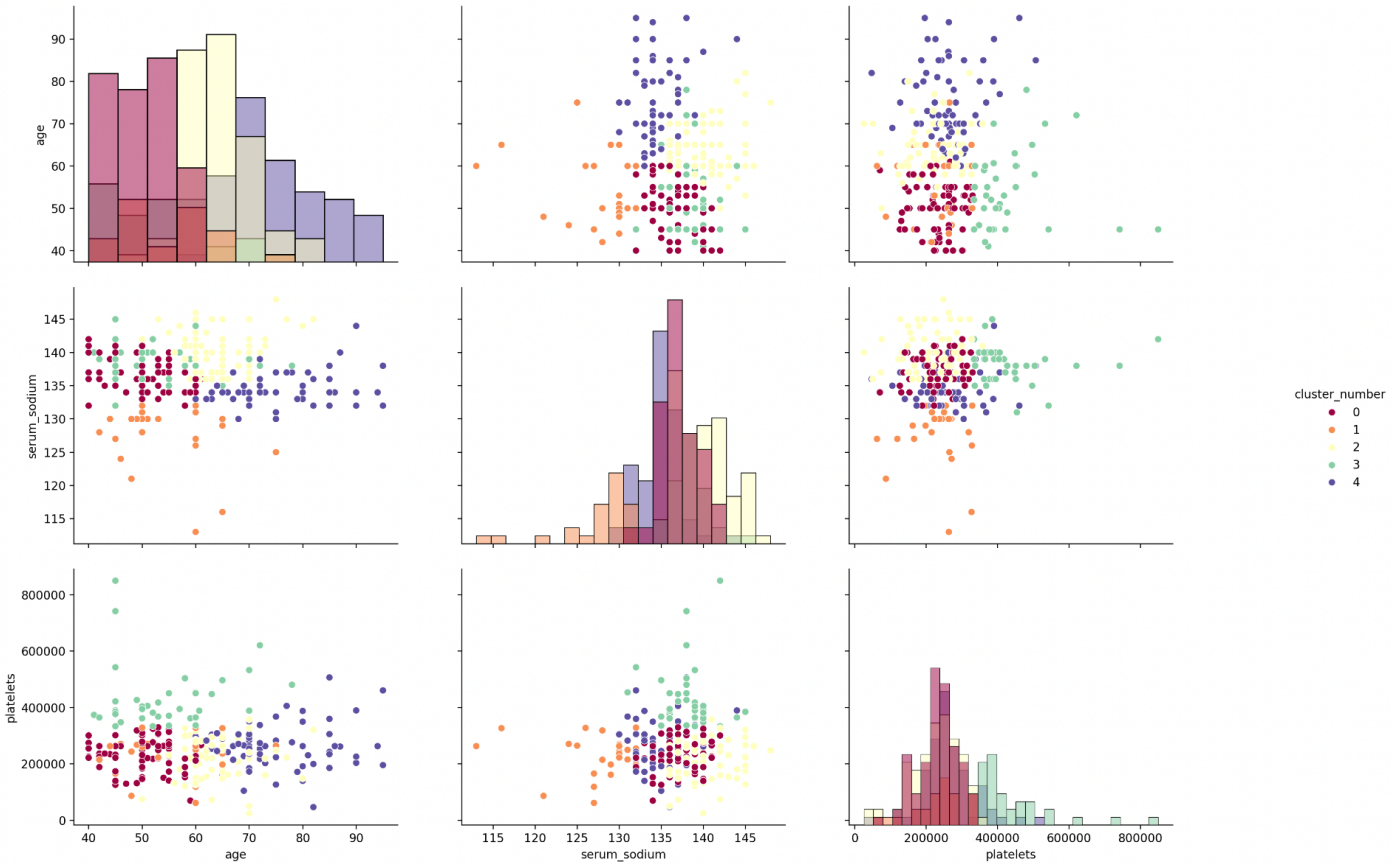
These are the value for each clusters:

Cluster Number	Total Records
0	84
2	83
4	69
3	38
1	25

Results

After clustering, we obtained these results:

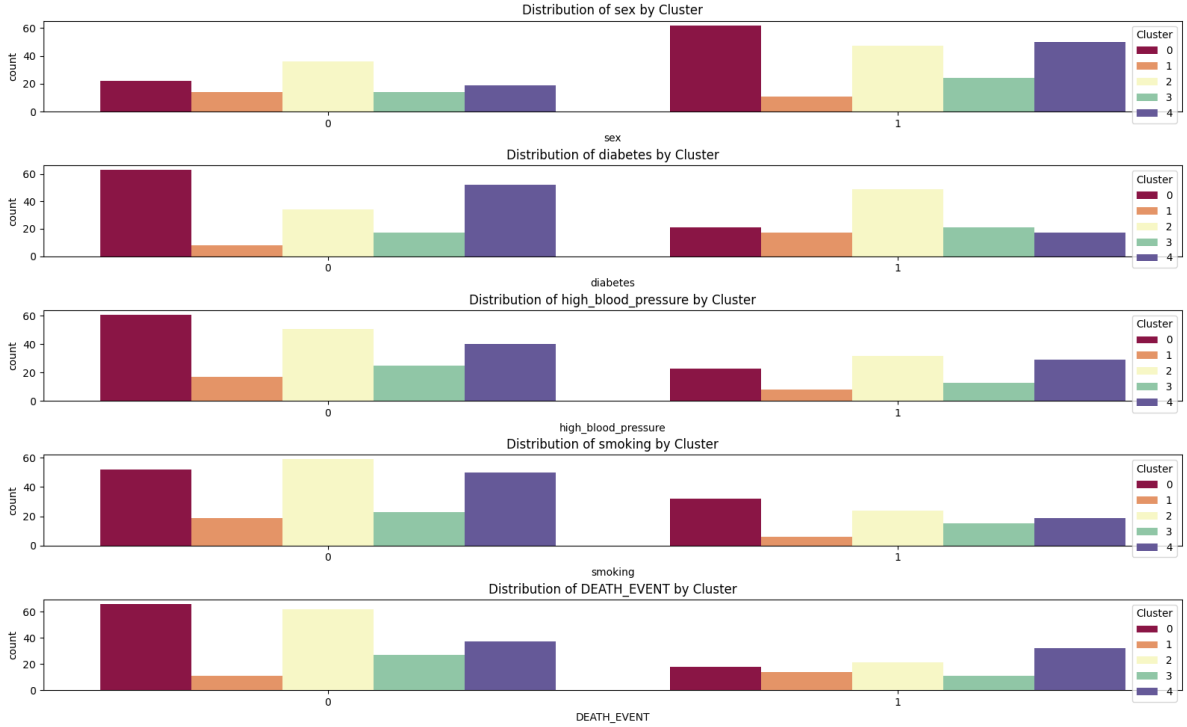
Numerical attributes after clustering



After clustering, we obtained five different group of clusters. We can see patient with younger age falls in the red, orange, and green cluster, while most older age falls in the purple and yellow clusters. Group with younger age tend to have lower serum sodium than older age, while for platelets, younger group has higher platelet count than older group.

Based on these observations, we can say that clustering has effectively identified different age, serum sodium, and platelet populations. This could potentially provide insights into various health conditions or demographic patterns.

Categorical attributes after clustering



Statistics across the 5 clusters:

Cluster 0:
Average Age: ~50.35 years
Sex: Predominantly male (73.8%)
Diabetes: 25% prevalence
High Blood Pressure: 27.4% prevalence
Serum Sodium: ~137.0 mEq/L
Platelets: ~233,316.43 kiloplatelets/mL
Smoking: 38.1% prevalence
Death Event: 21.4% prevalence

Cluster 1:
Average Age: ~55.24 years
Sex: Predominantly male (56.6%)
Diabetes: 68% prevalence
High Blood Pressure: 32% prevalence
Serum Sodium: ~127.52 mEq/L
Platelets: ~231,788.64 kiloplatelets/mL
Smoking: 24% prevalence
Death Event: 56% prevalence

Cluster 2:
Average Age: ~64.36 years
Sex: More balanced (56.6% male)
Diabetes: 59% prevalence
High Blood Pressure: 38.5% prevalence
Serum Sodium: ~140.01 mEq/L
Platelets: ~226,596.77 kiloplatelets/mL
Smoking: 28.9% prevalence
Death Event: 25.3% prevalence

Cluster 3:
Average Age: ~54.12 years
Sex: Predominantly male (63.2%)
Diabetes: 55.3% prevalence
High Blood Pressure: 34.2% prevalence
Serum Sodium: ~138.16 mEq/L
Platelets: ~430,842.11 kiloplatelets/mL
Smoking: 39.5% prevalence
Death Event: 28.9% prevalence

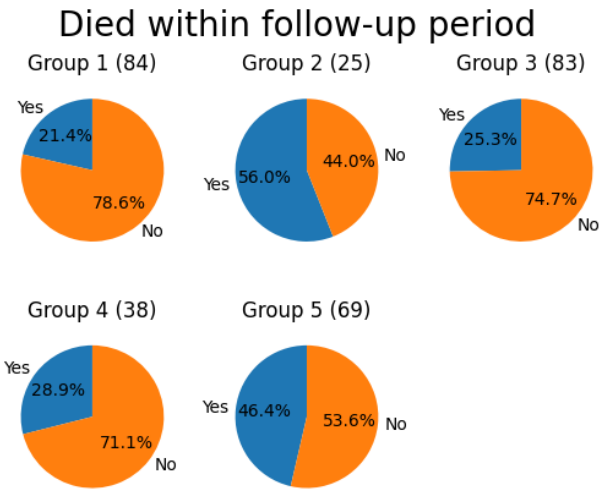
Cluster 4:
Average Age: ~75.09 years
Sex: Predominantly male (72.5%)
Diabetes: 24.6% prevalence
High Blood Pressure: 42% prevalence
Serum Sodium: ~134.57 mEq/L
Platelets: ~263,351.05 kiloplatelets/mL
Smoking: 27.5% prevalence
Death Event: 46.4% prevalence

What can we find from these 5 clusters statistic results?

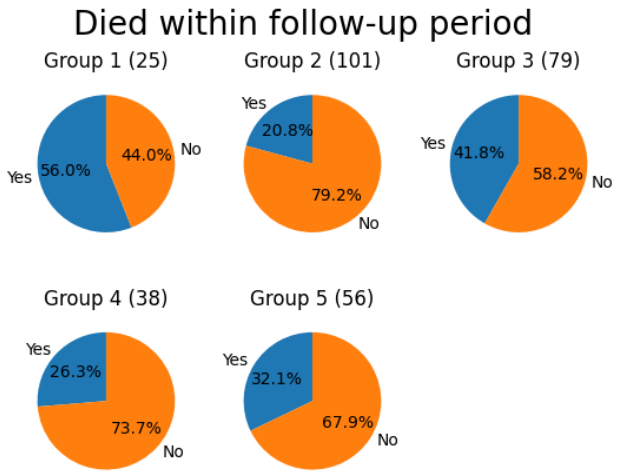
1. Distribution of age is clustered. Younger people tend to be in clusters 0, 1, and 3, while older people are more prevalent in clusters 2 and 4.
2. The distribution of sex is clustered, with more male-dominated clusters.
3. Diabetes, high blood pressure, and smoking also have distinct clustering patterns. Diabetes and high blood pressure tend to be more prevalent in older age groups, clusters 2 and 4,
4. Smoking is more common in younger age groups, clusters 0, 1, and 3.
5. Death event is high in cluster 1 and 4, and low in cluster 0, 2, and 3.

Finally, I'd like to note that I attempted to exclude the 'DEATH_EVENT' variable before performing the clustering analysis to assess whether there would be any discernible differences compared to when we include 'DEATH_EVENT' in the clustering. Here are the results:

Include DEATH_EVENT forclustering

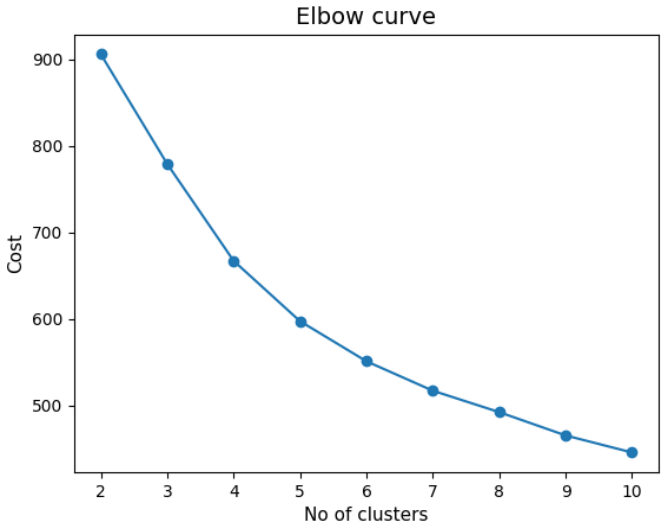


Without DEATH_EVENT for clustering



Indeed, the results indicate substantial similarities across certain clusters. However, it appears that the inclusion or exclusion of the 'DEATH_EVENT' variable might have led to minor adjustments in the assignment of specific records to clusters.

Without DEATH_EVENT for elbow method



It's also interesting to note that despite the minor changes observed in the elbow curve when excluding the 'DEATH_EVENT' attribute, the overall shape of the elbow curve remains quite similar, and the cost with slightly lower by a little. Moreover, the optimal number of clusters still appears to be 5. This suggests that the 'DEATH_EVENT' variable may not significantly alter the fundamental structure identified by the clustering algorithm.

Conclusions

To sum up, the methodical selection of an extensive dataset that includes both categorical and numerical attributes, along with the calculated implementation of the K-Prototypes Clustering algorithm, have combined to reveal important information about the complex terrain of heart failure. A sophisticated comprehension of the heterogeneous character of the condition has been produced by the algorithm's deft handling of a variety of features and its prudent application of statistical techniques for cluster identification. Heart failure subgroups are finely detailed by the clusters that have been identified and are distinguished by specific attributes. This increased clarity could lead to more individualised and focused treatment plans, which would benefit patients' outcomes and the management of heart failure as a whole.

References

[1] Heart failure clinical records. UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records

[2] Sklearn.preprocessing.StandardScaler. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[3] kprototype documentation. Developer Interface - kprototypes 0.1.2 documentation. (n.d.). https://kprototypes.readthedocs.io/en/latest/api.html

[4] Ruberts, A. (2020, May 16). *K-prototypes - customer clustering with mixed data types*. Well Enough. https://antonsruberts.github.io/kproto-audience/

[5] Suresh, A. (2020, December 1). *How to remove outliers for machine learning?*. Medium.c https://medium.com/analytics-vidhya/how-to-remove-outliers-for-machine-learning-24620c4657e8