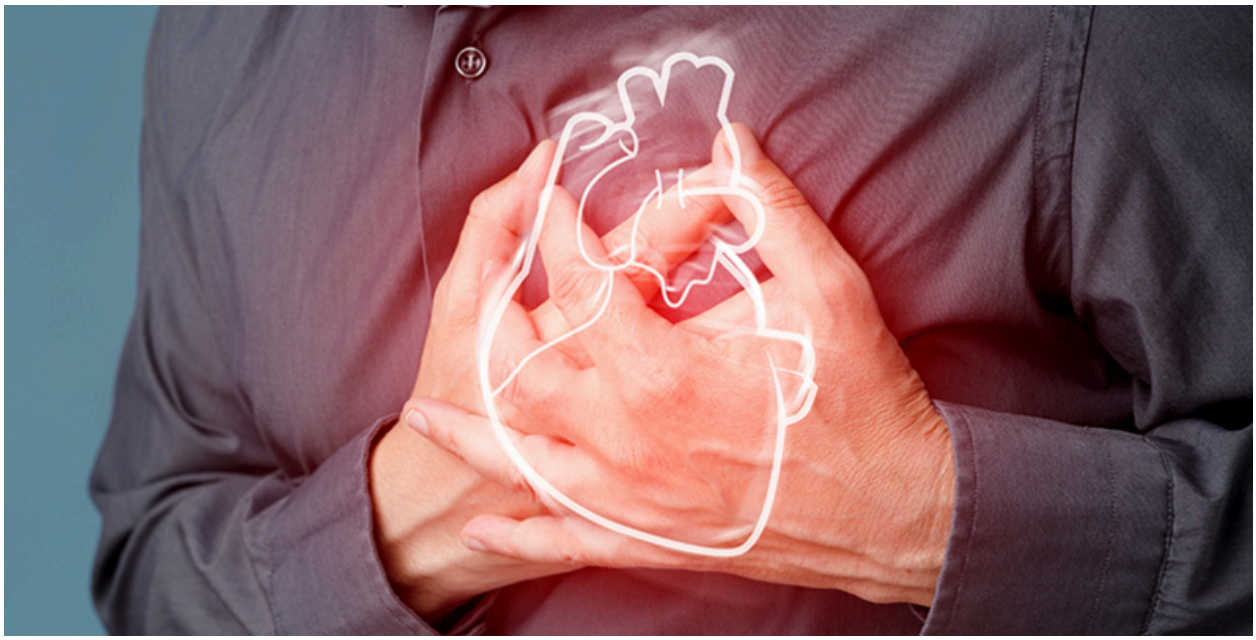


Predict survival of patients with heart failure

5/5/2024

Alex Kazmirchuk
David Iniguez
Nhan Nguyen

Overview

This project revolved around using Principal Component Analysis to explore a data set containing patients with heart failure. Our dataset was gathered at the Faisalabad Institute of Cardiology and Allied Hospital in Faisalabad, Punjab, Pakistan. Our data contains information regarding 299 heart failure patients of which 203 survived the trials where 96 unfortunately passed away. We are given feature data on each individual such as age, blood pressure, anemia, and gender; with the ultimate goal of our analysis being to find relationships with a death event and the other data using clustering algorithms and principal component analysis. To achieve this, we perform a principal component analysis on the dataset as a whole and sort subjects into different clusters. After our clusters are formed we hope to be able to discover patterns or relationships within the clusters that give insight into a patient dying before their next follow-up after already suffering from a heart failure.

The dataset contains 13 clinical features including age, anemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, time, and death event. With these features we are able to accurately represent an individual's health portfolio and make inferences about how long they may be able to live. In our analysis we determine some of these variables are stronger representations of either survival or passing away. We also determine which of these variables have the largest impact on the eventual outcome of the patient.

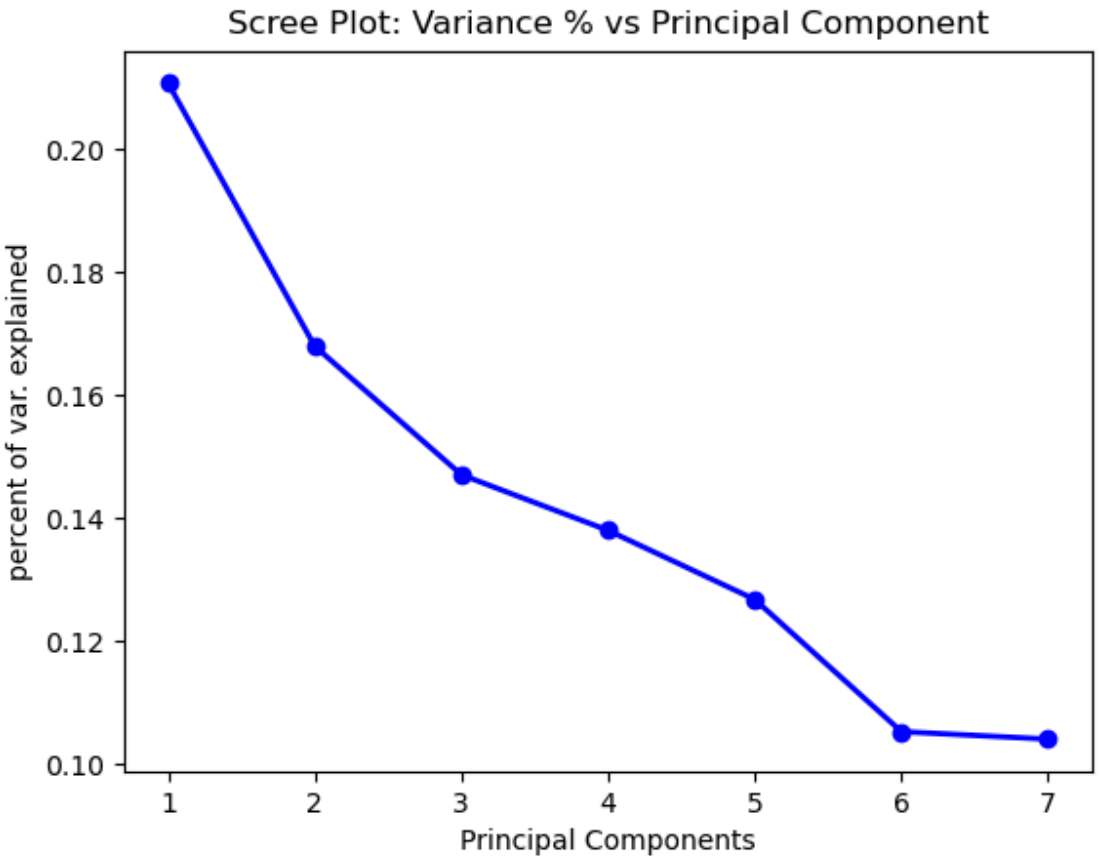
PCA Implementation

We'll utilize PCA to identify the principal axes of variation within our dataset. Our goal is to simplify the complex relationship among multiple clinical variables by reducing the

dimensionality of the data. PCA allows us to focus on the most significant features that contribute to patient outcomes.

Data preparation

Prior to performing PCA, the dataset was standardized to ensure each variable contributes equally. This involves scaling the data so that each feature has a mean of zero and a standard deviation of one. PCA was applied to only the seven standardized numeric features with mean zero and unit variance such as age, creatinine phosphokinases, ejection fraction, serum creatinine, and other standardized numeric data.



The scree plot shows a sharp decline in the variance explained from the first to the third and fifth principal components. This suggests that the first few components capture the most significant variance within the dataset. The elbow at the third and sixth components suggests that these components could be sufficient to capture the most crucial aspects of the data without significant information loss. Looking at the magnitude of the principal component transformation coefficients with magnitude greater than or equal to 0.3 highlighted in yellow

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
age	0.464962	0.452132	-0.007800	-0.198092	-0.191214	-0.634138	-0.318422
creatinine_phosphokinase	-0.137959	-0.193893	0.815054	-0.334406	0.294822	-0.100879	-0.264833
ejection_fraction	-0.178892	0.681478	-0.106713	-0.012995	0.469486	0.391348	-0.344178
platelets	-0.199258	0.246786	0.403317	0.820954	-0.180756	-0.173305	-0.007459
serum_creatinine	0.511777	0.045696	0.101672	0.182265	0.633580	-0.106913	0.528757
serum_sodium	-0.447411	0.429720	0.117976	-0.362607	-0.151399	-0.186519	0.641912
time	-0.480603	-0.214286	-0.370565	0.100469	0.446186	-0.598569	-0.135358

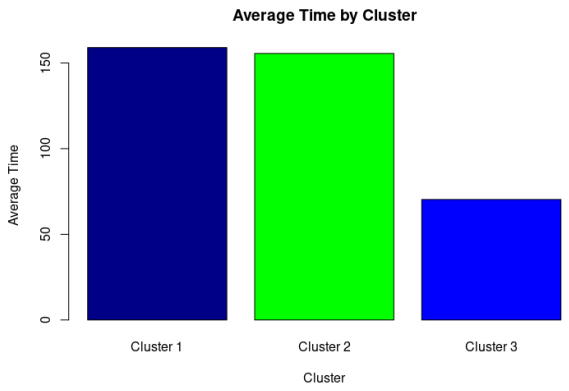
We can infer that age tends to be an important factor within 4 of the 7 principal components, and if we choose to keep the first three components based on the scree plot analysis we see that two out of the first three principal components have age as an important factor for the principal component (PC) transformation. Note also that within these first 2 PC's serum sodium levels are also significant factors. From these we can surmise that age and serum sodium levels will be important factors to keep in mind later during the cluster analysis.

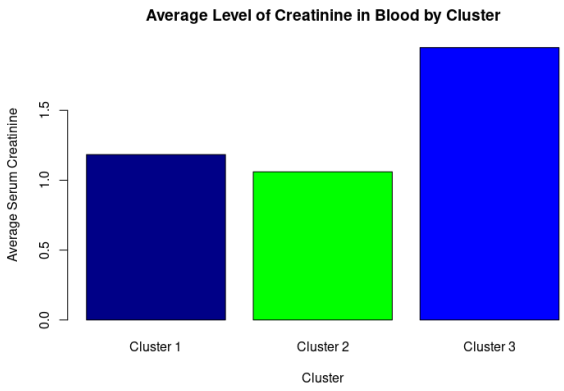
Cluster Analysis:

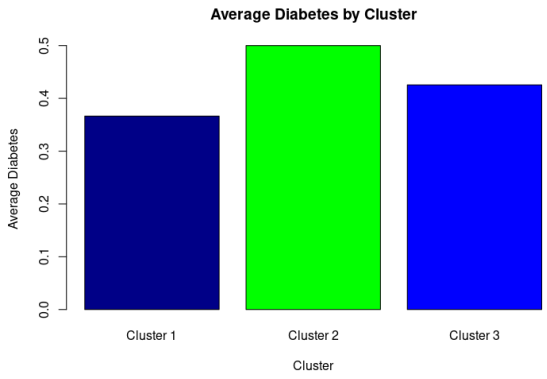
We start our analysis with three clusters. Our group realized that whether or not a person smoked was a very important variable for being sorted into a cluster, but was not a strong indication of whether they survived. From the scree plot analysis we felt three clusters were the best fit for our data. Using K=3, we began our cluster analysis by looking at the frequency of

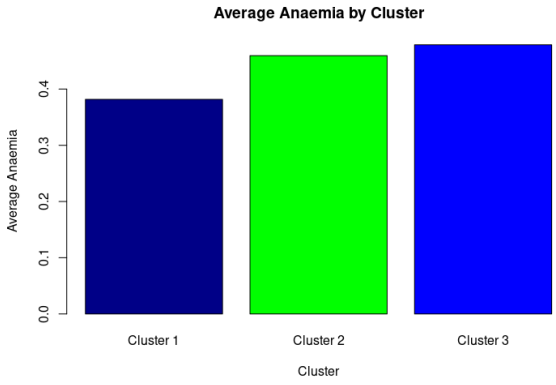
death events within each cluster to determine if there are key characteristics associated with a death happening after the patients were treated for heart failure.

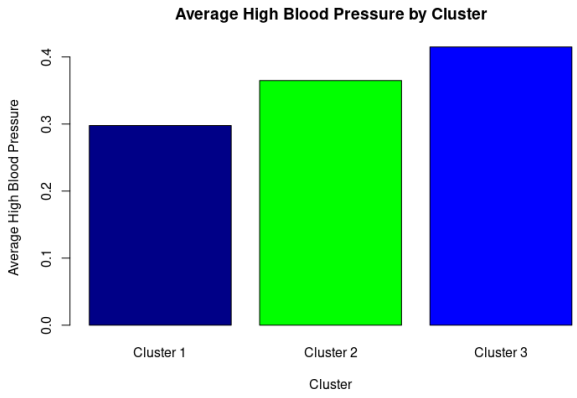
K = 3

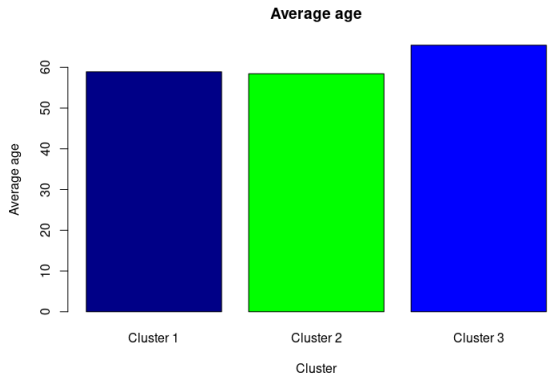












	Cluster	Death_Percentage
1	1	0.7633588
2	2	4.0540541
3	3	97.8723404

	Var1	Freq
1	1	131
2	2	74
3	3	94

	clust_b	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	time	DEATH_EVENT
1	1	58.90076	0.3816794	591.3969	0.3664122	39.28244	0.2977099	250173.1	1.183664	137.1221	1.00000000	158.98473	0.007633588
2	2	58.41442	0.4594595	464.1892	0.5000000	42.60811	0.3648649	295726.9	1.060405	137.5946	0.01351351	155.55405	0.040540541
3	3	65.43263	0.4787234	661.1383	0.4255319	32.85106	0.4148936	256250.9	1.949362	135.1702	0.65957447	70.31915	0.978723404

We defined our clusters as the following.

1. The healthiest group. This group had the lowest blood pressure out of the three groups as well as the lowest average for diabetes as well as anaemia. These characteristics are evident in the 99.2% survival rate for the group with only one of the 131 members passing away during the trial.
2. This group was also fairly healthy with significantly lower levels of Creatinine phosphokinase in the blood. They had the median blood pressure out of the three, but surprisingly the highest rate of diabetes. Despite this they had a 96% survival rate.
3. The most likely to die group. This group had significantly higher levels of Creatinine phosphokinase in the blood than any other group. They also had a higher average age than other groups and the highest blood pressure. Of the 94 members included in this group, 92 unfortunately passed away.

Within each cluster we wondered which principal component transformation had the largest (positive) values:

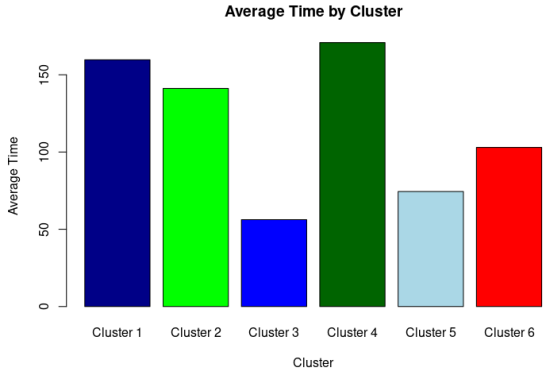
4. Within Cluster 1, 74% of the largest PC values came from PC2.
5. Within Cluster 2, 70% of the highest PC values came from PC3.
6. Within Cluster 3, 88% of the highest PC values came from PC1.

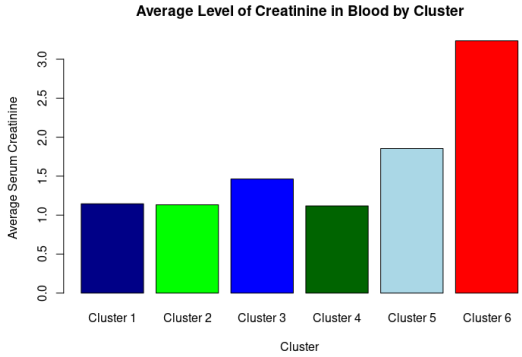
Looking back at PC1 we see that those patients with large PC1 values were older and had higher levels of serum creatinine and serum sodium. This indicates to us that these feature and patient metrics are important considerations for providing care to a patient after being treated for heart

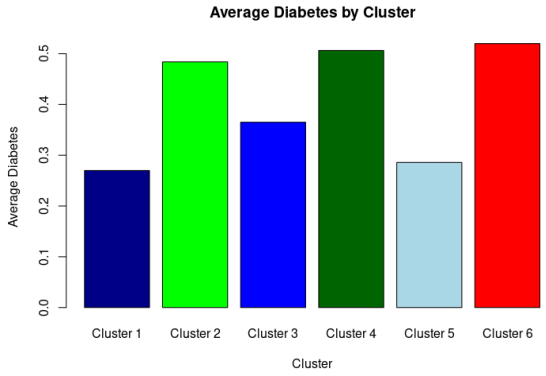
failure. Although this analysis yielded useful results, we felt there was little to distinguish which factors had an influence on healthy patients or which underlying causes may be leading to a patient's death. For this reason, we decided to increase the number of clusters to 6 in order to gain more insight into the factors influencing a patient's survival or death after heart failure. The graphs and analysis of the six clusters are on the next pages.

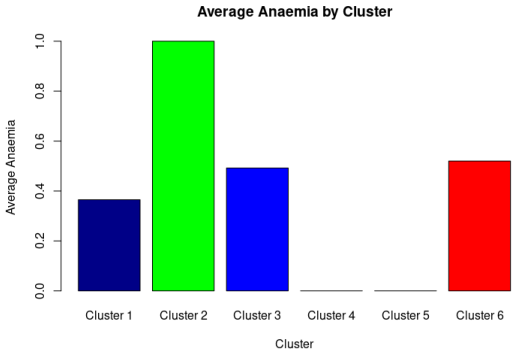
Detailed Comparison of Groups with Initial Model

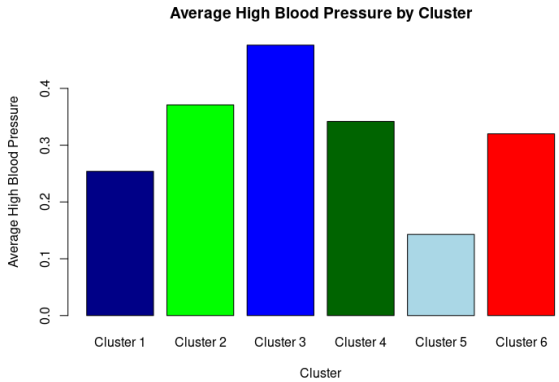
K = 6

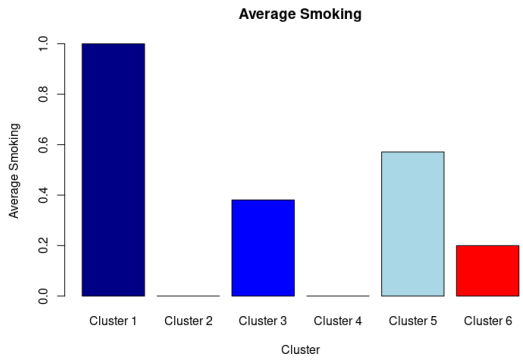












	Cluster	Death_Percentage
1	1	3.174603
2	2	3.225806
3	3	100.000000
4	4	5.063291
5	5	57.142857
6	6	84.000000

	Var1	Freq
1	1	63
2	2	62
3	3	63
4	4	79
5	5	7
6	6	25

We decided to define our clusters as the following.

1. Smokers with average to low health related issues. This group has low anemia, average high blood pressure, and overall no specific area where they have serious health issues. 97% of this cluster survived.
2. Non-smokers with high anemia and high diabetes. This group is interesting as they are a very polarizing group. They have 0 smokers, but several other high health related issues but this group also had a 97% survival rate.
3. High blood pressure who passes away quickly. This group has by far the largest problem with high blood pressure but also is about or above the average in every other category. This group has a 0% survival rate with all 63 members passing away.
4. Non-smokers with no anaemia. This group is objectively the healthiest out of any group with 0 members smoking as well as 0 members with anaemia. They have a 95% survival rate with the greatest frequency of members.
5. Non-anaemia with average health related issues. This group is pretty average compared to the other clusters and only has 7 members. Of the 7 members 3 survived and 4 passed away.

6. High levels of creatinine. This group displays extremely high values of creatinine compared to other groups. They also had a large value for diabetes and had a low survival rate at only 16%.

Conclusion

During our analysis we were able to confidently sort all subjects in our data into six groups. After performing our principal component analysis on our numeric feature variables, we get meaningful insight into which features are playing an important role in this analysis. Clusters 3 and 6 have a very high mortality rate in contrast to other clusters. There are more patients within cluster 3 that have been diagnosed with high blood pressure than any of the other clusters. The high mortality rate within this cluster may be due to the presence of high blood pressure. Within cluster 6, these patients had higher levels of serum creatinine and also had more diabetes diagnosis than any other cluster, indicating that deaths in this cluster are primarily due to those features. Of significant importance are the three clusters with lower mortality rates; cluster 1, 2, and 4. A common factor among the three clusters with low mortality rates is the low average serum creatinine levels in these clusters indicating that low levels of this health metric indicates a 'healthier' patient. Clusters 2 and 4 tend to have high diabetes diagnosis, when we consider this in conjunction with the lower mortality rate it may indicate that diabetes may not be indicative of death after heart failure; which may require more investigation beyond the scope of this paper to determine truthfulness of that hypothesis. Cluster 2 and 4 have zero smokers, but cluster 1 has 100% of the patients in that cluster smoke. This indicates that smoking is also not a strong factor for identifying people at risk of death after heart failure. Regardless, the smoking variable can be

useful in creating distinct clusters in our data which we concluded after our discussion last class period.

These results can be useful for physicians as they can assess patients based on their preexisting conditions and properly prepare them for the future. If a patient is expected to pass away they can assess the time each patient may have less and inform them of how much time they have to spend with family. If precautions may be taken, more lives may be able to be saved if patients are properly prepared for the time post heart failure.