Analysis of Health Data Associated with Heart Failure

Name	Email	Contributions
Lillian Quynn	ljquynn@ucdavis.edu	PCA, plots, interpretation
Marco Oviedo	jmoviedo@ucdavis.edu	K-means clustering, plots, function development
Melanie Bluck	mblu@ucdavis.edu	Report writing, prediction model
Rishit Puri	rispuri@ucdavis.edu	Logistic regression
Xinran Zhou	axrzhou@ucdavis.edu	K-means clustering

Table of Contents

Introduction

Cardiovascular diseases account for a plurality of all deaths globally, and heart failure is a devastating outcome of poor cardiovascular health. High-risk patients or those who have already been diagnosed with a cardiovascular condition need preventative care and early detection to avoid a fatality by heart failure. This is a need that can be aided by machine learning models which can quantify the risk possessed by a wide variety of patients. In this report, we will analyze a variety of health data obtained from patients who have suffered heart failure and build a model to predict fatality.

Dataset Description

The heart failure dataset was obtained from Kaggle.com. It is a recent dataset from 2020. The dataset contains 299 observations. There are 13 variables, 5 of which are binary categorical variables and 7 are numeric variables. The numeric variables describe various blood components obtained from a blood sample. The variables are defined as follows:

- Age – The age of the patient
- Anaemia – If the patient is anaemic or not
- Creatinine phosphokinase – Level of the CPK enzyme in the blood (mcg/L)
- Diabetes – If the patient is diabetic or not
- Ejection fraction – Percentage of blood leaving the heart in each contraction
- High blood pressure – If the patient has hypertension or not
- Platelets – Level of platelets in the blood (kiloplatelets/mL)
- Serum creatinine – Level of serum creatinine in the blood (mg/dL)
- Serum sodium – Level of serium sodium in the blood (mEq/L)
- Sex – If the patient is male or female
- Smoking – If the patient smokes or not
- Time – Follow-up period (days)
- Death event – If the patient was deceased during the follow-up period

We will use only the numeric variables in the unsupervised learning except for sex and smoking, which will be used to group the data. We will use all variables in the supervised learning. The dataset has already been cleaned.

Research Questions

- What general insights can be drawn from the blood components? Which of them are correlated, and how much of a relationship exists between them?
- Do the relationships between blood components differ among males and females?
- How does smoking impact blood components?
- How accurately can the variables in our data predict the chances of fatality from heart failure?

Unsupervised Learning

PCA

First, we built a correlation matrix to see if there are any correlations that exist between the five blood components.

Figure 1: Correlation Matrix

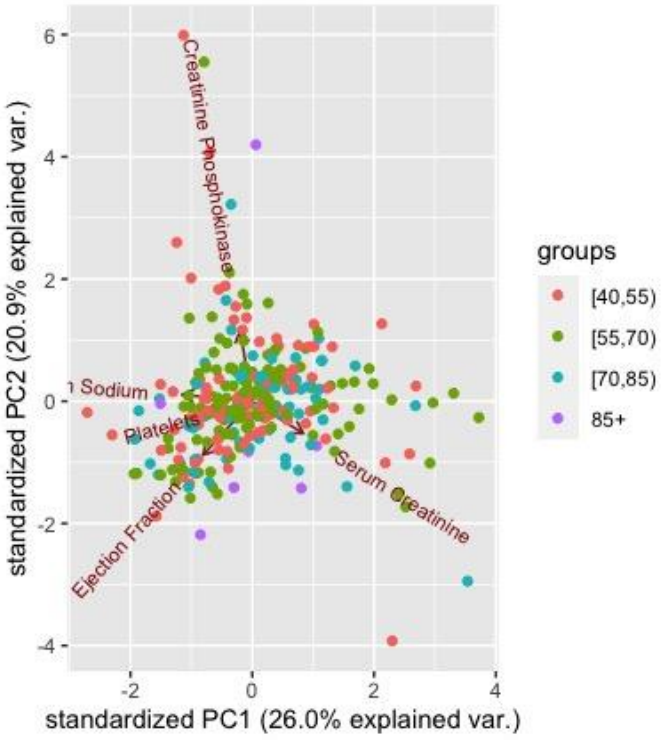
	Creatinine Phosphokinase	Serum Creatinine	Serum Sodium	Platelets	Ejection Fraction
Creatinine Phosphokinase	1	-0.016	0.059	0.0245	-0.044
Serum Creatinine	-0.016	1	-0.189	-0.041	-0.011
Serum Sodium	0.059	-0.189	1	0.062	0.176
Platelets	0.0245	-0.041	0.062	1	0.072
Ejection Fraction	-0.044	-0.011	0.176	0.072	1

The greatest correlation coefficient is -0.189 between serum creatinine and serum sodium. Even this is a very weak association, and so it appears that no meaningful correlations exist between any of the blood components. We proceed to scale the data and generate principle components, whose loadings are displayed in Figures 2 and 3 below.

Figure 2: Loadings Table

	PC1	PC2	PC3	PC4	PC5
Creatinine Phosphokinase	-0.1149	0.7437	-0.4362	-0.4441	0.2148
Serum Creatinine	0.4816	-0.3385	-0.4875	-0.4405	-0.4710
Serum Sodium	-0.6607	0.0679	0.1677	-0.2162	-0.6956
Platelets	-0.3219	-0.1003	-0.7253	0.5967	-0.0644
Ejection Fraction	-0.4633	-0.5636	-0.13336	-0.4538	0.4939

Figure 3: Biplot



For additional insight, the data in the biplot have been color-coded according to age range. No clustering occurs in the biplot, and the different ages seem to be scattered randomly. The plot also shows that PC1 and PC2 together explain 46.9% of the variation in the data, which is quite low. Our PCA analysis suggests that there are few if any patterns to be observed among the blood components represented in the dataset.

K-Means Clustering

We know introduce two of our categorical variables into the analysis – smoking and sex. We chose k-means clustering as the unsupervised clustering method best suited to our binary variables. We scaled the data and used the principle components obtained from the PCA to reduce the dimensions of the data so that the clustering results can be displayed in the following two plots.

Figure 4: Sex Cluster Plot

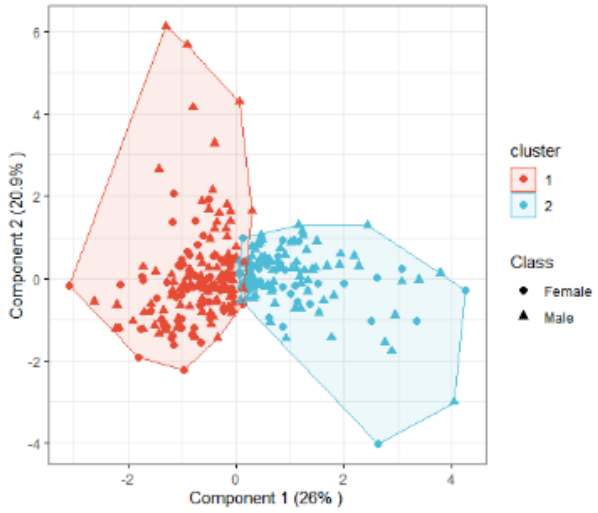
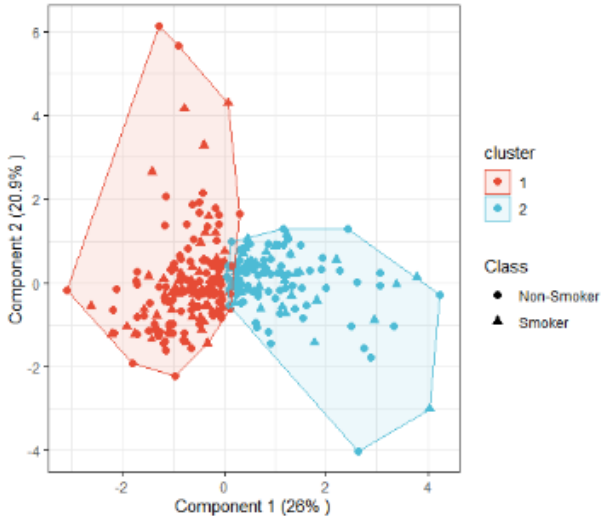


Figure 5: Smoking Cluster Plot



Neither the classes for sex nor smoking show any distinction between the clusters, and no pattern is revealed here. It seems that sex and smoking have no relationship to the blood components in our dataset, which reinforces the lack of patterns found in the PCA.

Supervised Learning

Logistic Regression

To build our predictive model for fatalities, we begin by fitting all the variables to a logistic regression model and obtaining coefficient estimates, as displayed below.

Figure 6: Model Summary (All Variables)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	10.1849290225454	5.65657032464257	1.80054846629861	0.0717740779628756
Age	0.047419073553999	0.0158006323015473	3.00108708620197	0.00269017615766706
Anemia	-0.00747045170792	0.360489086258031	-0.02072310090014	0.983466541122091
C. Phospho-kinase	0.000222229382816	0.000177933244262	1.24894807452994	0.211684066460838
Diabetes	0.145149775258558	0.351188640232869	0.413309995341281	0.679379507619881
Ejection Fraction	-0.07666250137550	0.0163291296591868	-4.69483083149942	2.6682744835874e-06
High Blood Pressure	-0.10267942700761	0.358706892616997	-0.28624882632866	0.774687549524315
Platelets	-1.1996244604e-06	1.889060343884e-06	-0.63503766003666	0.525403853411978
Serum Creatinine	0.666093339634976	0.181492575641695	3.67008588246599	0.0002424689903281
Serum Sodium	-0.06698107216303	0.0397350980902629	-1.68569036902527	0.0918554528607754
Sex	-0.53365801594898	0.41391803903357	-1.28928426795554	0.197299278217744
Smoking	-0.01349222424478	0.412617797959162	-0.03269908450754	0.973914553959325
Time	-0.02104462583502	0.00301439395076304	-6.98137873773714	2.922971401061e-12

Most coefficients are too small to be meaningful. Among them, serum creatinine (0.6661) and sex (-0.5337) are the greatest, but sex is also associated with a p-value of 0.1973, well above the standard 0.05 limit. In contrast, serum creatinine has a p-value of 0.0002, which is highly significant. Therefore, our final model will include serum creatinine. While ejection fraction and time have small coefficients (-0.0210 and -0.0766, respectively), they're highly significant with p-values close to zero (approx. 2.92e-12 and 2.67e-06). Age also has a small coefficient of 0.047 but high significance (p = 0.0027). Despite

Model Fitting

We now use stepwise selection to determine which variables are in our final model. The model with the smallest AIC is summarized below. It contains the same variables discussed above.

Figure 7: Model Summary (Min. AIC)

	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	9.493034	5.405768	1.756	0.07907
Age	0.042466	0.015030	2.825	0.00472
Ejection fraction	-0.073430	0.015785	-4.652	3.29e-06
Serum creatinine	0.685990	0.174044	3.941	8.10e-05
Serum sodium	-0.064557	0.038377	-1.682	0.09254
Time	-0.020895	0.002916	-7.166	7.74e-13

Prediction

We split the data into training and testing sets, where the testing set is the first 20 rows with a fatality and the first 20 rows without a fatality. After training and testing the model, we obtain the following confusion matrix.

Figure 8: Confusion Matrix

		TRUE	
PREDICTED		Not Fatal	Fatal
	Not Fatal	10	10
	Fatal	2	18

The model has an overall accuracy of 70%. It tends to overestimate the chance of fatality, as 83% of the errors are false positives.

Interpretation of Results

- What general insights can be drawn from the blood components? Which of them are correlated, and how much of a relationship exists between them?

Our PCA didn't show any patterns among the blood components, and no meaningful correlation was established between any of them. We have insufficient evidence to conclude that any such relationships exist.

- Do the relationships between blood components differ among males and females?
- How does smoking impact blood components?

The clustering analysis didn't show any relationship between sex and the values of the blood components, and likewise for smoking. Furthermore, neither sex nor smoking had any meaningful effect on the chance of dying after a diagnosis of heart failure, as shown by logistic regression. We don't have any evidence to conclude there exists relationships among the blood components or fatality between sex or smoking.

- How accurately can the variables in our data predict the chances of fatality from heart failure?

We were able to build a logistic regression model that predicted whether a patient would die with 70% accuracy. The false predictions mainly stem from overestimation of the chance of dying, meaning the chance of a false prediction is higher for patients who are forecasted to die than those forecasted to live.

Conclusion

While much of our analysis yielded little, we did gain valuable insight from our predictive model by showing which health metrics affected the chance of dying after a diagnosis of heart failure. The risks and needs of a patient with poor cardiovascular health can be determined by medical professionals using a blood sample and testing for sodium and creatinine levels. It also reinforces the importance of tests like the echocardiogram, which measure the heart's ejection fraction. We urge doctors to perform these tests often in high-risk patients. Additionally, we encourage doctors to extend their patients' follow-up periods after a diagnosis of heart failure, as this was shown to reduce the lethality.

References

DataCamp. (n.d.). *Eigenvalue: Extract and visualize the eigenvalues/variances of dimensions*. RDocumentation. Retrieved September 5, 2022, from https://www.rdocumentation.org/packages/factoextra/versions/1.0.7/topics/eigenvalue

DataCamp. (n.d.). *GET_PCA: Extract the results for individuals/variables - PCA*. RDocumentation. Retrieved September 5, 2022, from https://www.rdocumentation.org/packages/factoextra/versions/1.0.7/topics/get_pca

DataCamp. (n.d.). *GGSCATTER: Scatter plot*. RDocumentation. Retrieved September 4, 2022, from https://www.rdocumentation.org/packages/ggpubr/versions/0.4.0/topics/ggscatter

Davide Chicco, Giuseppe Jurman: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making 20, 16 (2020).

Larxel. (2020, June 20). *Heart failure prediction*. Kaggle. Retrieved August 29, 2022, from https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data?resource=download

Mankad, R. (2021, February 26). *Ejection fraction: An important heart test*. Mayo Clinic. Retrieved September 6, 2022, from https://www.mayoclinic.org/tests-procedures/ekg/expert-answers/ejection-fraction/faq-20058286

Appendix: R Code

```
library(dplyr)
library(ggbiplot)
library(knitr)
library(kableExtra)
library(ggpubr)
library(janitor)
library(MASS)
library(ggplot2)
library(tidyverse)
library(caret)
library(factoextra)
```

PCA

```
data=heart_failure_clinical_records_dataset
library("dplyr")
age_group <- data %>% mutate(case_when(data$age >= 85 ~ '85+',
                                       data$age >= 70  & age < 85 ~ '[70,85)',
                                       data$age >= 55  & age < 70 ~ '[55,70)',
                                       data$age >= 40  & age < 55 ~ '[40,55)')) # end function
age_group

df=data.frame(data$creatinine_phosphokinase, data$serum_creatinine,

              data$serum_sodium, data$platelets, data$ejection_fraction)
df

new_data <- df %>% rename("Platelets" = "data.platelets",
                          "Serum Creatinine"="data.serum_creatinine",
                          "Serum Sodium"="data.serum_sodium",
                          "Ejection Fraction"="data.ejection_fraction",
                          "Creatinine Phosphokinase"="data.creatinine_phosphokinase")
new_data
cor(df)
datapca <- prcomp(new_data,sc=TRUE)
datapca
summary(datapca)
ggbiplot(datapca,  groups=age_group$`case_when(...)`)
screeplot(datapca, type="lines")
```

CLUSTERING

```
Heart <- heart_failure_clinical_records_dataset
Heart$smoking [Heart$smoking == '1'] <- "Smoker"
Heart$smoking [Heart$smoking == '0'] <- "Non-Smoker"
km <- kmeans(scale(Heart[, c(3,5,7:9)]), 2, nstart = 25)
km
```

ORIGINAL FUNCTION

```
cluster <- function(data, cat){
  pca <- prcomp(data,  scale = TRUE)
  Cluster.coord <- as.data.frame(get_pca_ind(pca)$coord)
  Cluster.coord$cluster <- factor(km$cluster)
  Cluster.coord$Class <- cat
  eigenvalue <- round(get_eigenvalue(pca), 1)
  variance.percent <- eigenvalue$variance.percent
  x <- table(Cluster.coord$Class,Cluster.coord$cluster)/length(Cluster.coord$Class)
  plot <- ggscatter(
  Cluster.coord, x = "Dim.1", y = "Dim.2", main = "Cluster Plot",
  color = "cluster", palette = "npg", ellipse = TRUE, ellipse.type = "convex",
     shape = "Class", size = 2.25,  legend = "right", ggtheme = theme_bw(),
  xlab = paste0("Component 1 (", variance.percent[1], "% )" ),
  ylab = paste0("Component 2 (", variance.percent[2], "% )" )
)
  mylist <- list(x, plot)
  return(mylist)
}
```

```
cluster(Heart[, c(3,5,7:9)], Heart[, 11])          ##Smoking plot/Distribution table
```

```
Heart$sex [Heart$sex == '1'] <- "Male"
Heart$sex [Heart$sex == '0'] <- "Female"
```

```
cluster(Heart[, c(3,5,7:9)], Heart[, 10])          ##Gender plot/Distribution table
```

LOGISTIC REGRESSION

```
data <- read.csv("heart_failure.csv")

full.model <- glm(DEATH_EVENT ~ ., data = data, family = "binomial")

summary(full.model)

step.model <- stepAIC(full.model, direction = "both", trace = F)

summary(step.model)

test.data <- rbind(data[data$DEATH_EVENT == 1,][1:20,], data[data$DEATH_EVENT == 0,][1:20,])

train.data <- rbind(data[data$DEATH_EVENT == 1,][21:96,], data[data$DEATH_EVENT == 0,][21:203,])

train.model <- glm(DEATH_EVENT ~ age + ejection_fraction + serum_creatinine + serum_sodium + time,
data = train.data, family = "binomial")

test.data$DEATH_EVENT <- factor(test.data$DEATH_EVENT)

pred <- factor(as.integer(ifelse(predict(train.model, newdata = test.data, type = "response")>.5, '1', '0'))

confusion <- confusionMatrix(test.data$DEATH_EVENT, pred)

confusion
```