Final Project: Heart Failure Prediction

Kolekar, Ajit

November 16th, 2020

## Overall, write a coherent narrative that tells a story with the data as you complete this section.

The most common cause of death globally is Cardiovascular Diseases. Approximately 18 million people die due to Cardiovascular Diseases every year, which is around 31% of all deaths. Heart failure is the most common event that can be caused by Cardiovascular Diseases. There are various factors that can contribute to Cardiovascular Diseases. Through this project, I analyzed the factors that contribute to Cardiovascular Diseases and attempted to predict the preventative measures that need to be taken to avoid heart attacks.

In order to achieve the objectives of the project, I used the dataset located at <https://www.kaggle.com/andrewmvd/heart-failure-clinical-data>.

heart\_df <- read.csv("C:/Users/Shilp/Documents/GitHub/dsc520/data/heart\_failure\_clinical\_records\_dataset.csv")  
str(heart\_df)

## 'data.frame': 299 obs. of 13 variables:  
## $ age : num 75 55 65 50 65 90 75 60 65 80 ...  
## $ anaemia : int 0 0 0 1 1 1 1 1 0 1 ...  
## $ creatinine\_phosphokinase: int 582 7861 146 111 160 47 246 315 157 123 ...  
## $ diabetes : int 0 0 0 0 1 0 0 1 0 0 ...  
## $ ejection\_fraction : int 20 38 20 20 20 40 15 60 65 35 ...  
## $ high\_blood\_pressure : int 1 0 0 0 0 1 0 0 0 1 ...  
## $ platelets : num 265000 263358 162000 210000 327000 ...  
## $ serum\_creatinine : num 1.9 1.1 1.3 1.9 2.7 2.1 1.2 1.1 1.5 9.4 ...  
## $ serum\_sodium : int 130 136 129 137 116 132 137 131 138 133 ...  
## $ sex : int 1 1 1 1 0 1 1 1 0 1 ...  
## $ smoking : int 0 0 1 0 0 1 0 1 0 1 ...  
## $ time : int 4 6 7 7 8 8 10 10 10 10 ...  
## $ DEATH\_EVENT : int 1 1 1 1 1 1 1 1 1 1 ...

The dataset includes data of 299 patients. It has following 13 columns and their description.

* age - Age of the patient (Numeric value)
* anaemia - Decrease of red blood cells or hemoglobin (Boolean value - 0 for No and 1 for Yes)
* creatinine\_phosphokinase - Level of the CPK enzyme in the blood (Integer value - mcg/L)
* diabetes - If the patient has diabetes (Boolean value - 0 for No and 1 for Yes)
* ejection\_fraction - Percentage of blood leaving the heart at each contraction (Integer value - percentage)
* high\_blood\_pressure - If the patient has hypertension (Boolean value - 0 for No and 1 for Yes)
* platelets - Platelets in the blood (Numeric value - kiloplatelets/mL)
* serum\_creatinine - Level of serum creatinine in the blood (Numeric value - mg/dL)
* serum\_sodium - Level of serum sodium in the blood (Integer value - mEq/L)
* sex - Woman or man (Boolean value - 0 for Female and 1 for Male)
* smoking - If the patient smokes or not (Boolean value - 0 for No and 1 for Yes)
* time - Follow-up period (Integer value - days)
* DEATH\_EVENT - If the patient deceased during the follow-up period (Boolean value - 0 for No and 1 for Yes)

DEATH\_EVENT is the dependent (outcome) variable while all other variables are independent variables.

I reviewed the data in the CSV file and determined that all the rows were fully populated. There were also no missing values. My review also indicated that the contents of the data did not have any outliers. I removed columns: platelets and serum\_sodium from the dataset. No other cleaning steps were needed for the dataset.

Following are the cleanup steps and a quick look at the data after I performed the cleanup activities.

heart\_df$platelets <- NULL  
heart\_df$serum\_sodium <- NULL  
head(heart\_df)

## age anaemia creatinine\_phosphokinase diabetes ejection\_fraction  
## 1 75 0 582 0 20  
## 2 55 0 7861 0 38  
## 3 65 0 146 0 20  
## 4 50 1 111 0 20  
## 5 65 1 160 1 20  
## 6 90 1 47 0 40  
## high\_blood\_pressure serum\_creatinine sex smoking time DEATH\_EVENT  
## 1 1 1.9 1 0 4 1  
## 2 0 1.1 1 0 6 1  
## 3 0 1.3 1 1 7 1  
## 4 0 1.9 1 0 7 1  
## 5 0 2.7 0 0 8 1  
## 6 1 2.1 1 1 8 1

## Summarize the problem statement you addressed.

The problem statement of the project was to determine the factors that could contribute to the death by heart failure using historical data and to predict how the death by heart failure can be avoided by taking necessary steps. In order to address the problem statement, I focused on following research questions.

1. Does diabetes contribute to death due to heart failure?
2. What impact does age have on the risk of death due to heart failure?
3. Does anemia contribute to death due to heart failure?
4. Does the amount of creatinine phosphokinase in heart contribute to heart failure?
5. How does ejection fraction relate to the heart’s health?
6. Does high blood pressure contribute to death due to heart failure?
7. What impact does gender have on the risk of death due to heart failure?
8. Could smoking increase the risk of death by heart failure?
9. What other factors could contribute to death due to heart failure?

## Summarize how you addressed this problem statement (the data used and the methodology employed).

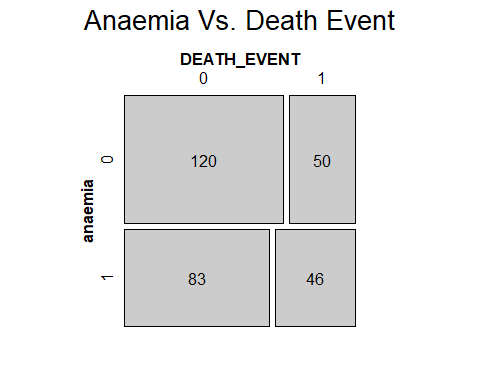
After cleaning the dataset, I conducted exploratory data analysis to understand how each variable impacts the DEATH\_EVENT variable. In order to do so, I created mosaic plots and box plots as below.

library('ggplot2')  
library('vcd')

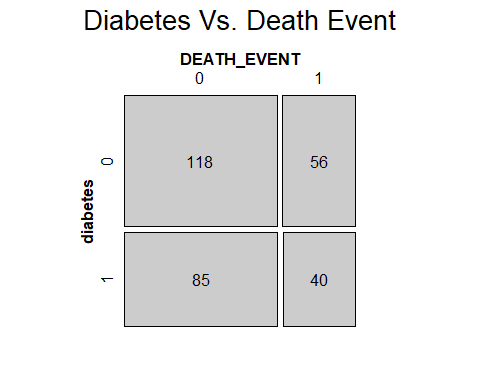
## Warning: package 'vcd' was built under R version 4.0.3

## Loading required package: grid

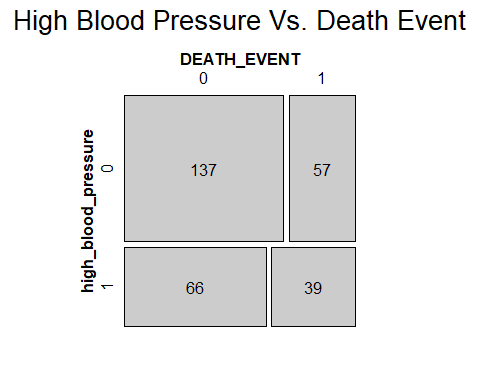
heart\_factor\_df <- heart\_df  
heart\_factor\_df$DEATH\_EVENT <- as.factor(heart\_factor\_df$DEATH\_EVENT)  
  
anaemia\_table <- structable(~ anaemia + DEATH\_EVENT, data=heart\_factor\_df)  
mosaic(anaemia\_table, main="Anaemia Vs. Death Event", pop=FALSE)  
labeling\_cells(text = as.table(anaemia\_table), margin = 0)(as.table(anaemia\_table))



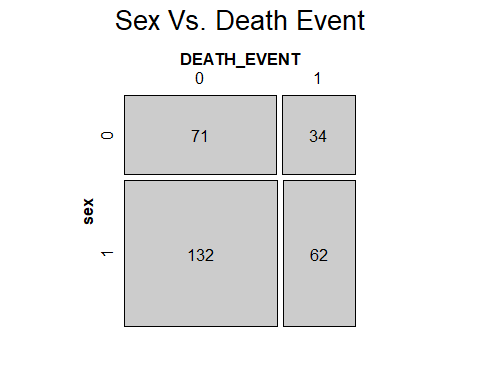
diabetes\_table <- structable(~ diabetes + DEATH\_EVENT, data=heart\_factor\_df)  
mosaic(diabetes\_table, main="Diabetes Vs. Death Event", pop=FALSE)  
labeling\_cells(text = as.table(diabetes\_table), margin = 0)(as.table(diabetes\_table))



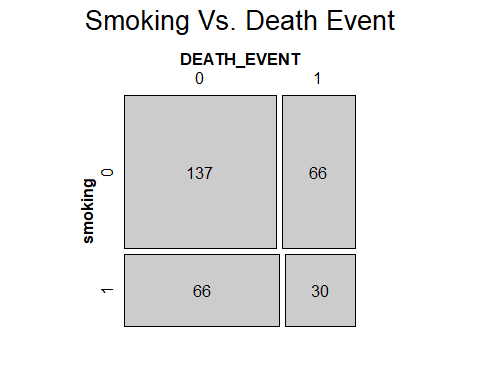
hbp\_table <- structable(~ high\_blood\_pressure + DEATH\_EVENT, data=heart\_factor\_df)  
mosaic(hbp\_table, main="High Blood Pressure Vs. Death Event", pop=FALSE)  
labeling\_cells(text = as.table(hbp\_table), margin = 0)(as.table(hbp\_table))



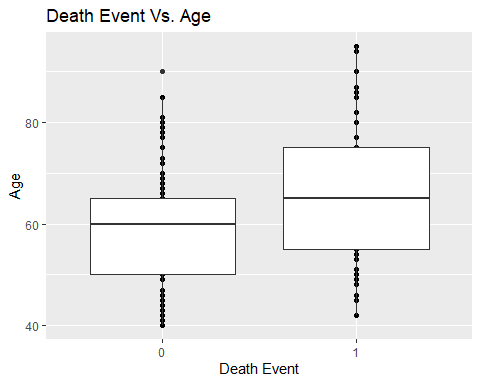
sex\_table <- structable(~ sex + DEATH\_EVENT, data=heart\_factor\_df)  
mosaic(sex\_table, main="Sex Vs. Death Event", pop=FALSE)  
labeling\_cells(text = as.table(sex\_table), margin = 0)(as.table(sex\_table))



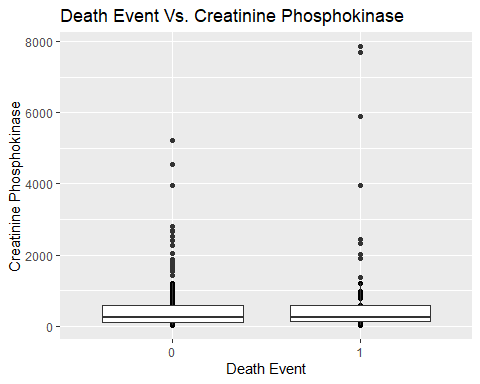
smoking\_table <- structable(~ smoking + DEATH\_EVENT, data=heart\_factor\_df)  
mosaic(smoking\_table, main="Smoking Vs. Death Event", pop=FALSE)  
labeling\_cells(text = as.table(smoking\_table), margin = 0)(as.table(smoking\_table))



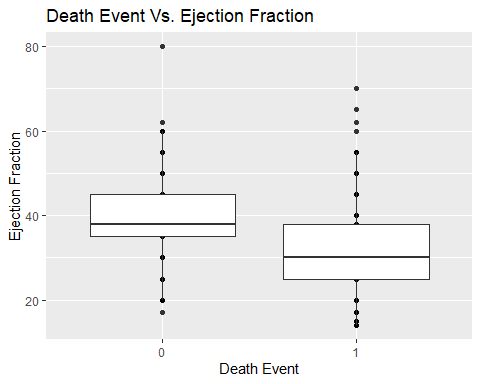
ggplot(heart\_factor\_df, aes(x=DEATH\_EVENT, y=age)) + geom\_point() + geom\_boxplot() + ggtitle("Death Event Vs. Age") + xlab("Death Event") + ylab("Age")



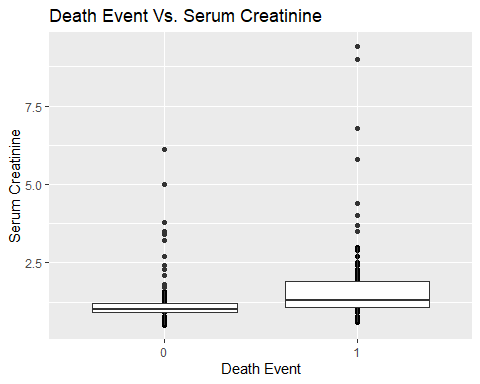
ggplot(heart\_factor\_df, aes(x=DEATH\_EVENT, y=creatinine\_phosphokinase)) + geom\_point() + geom\_boxplot() + ggtitle("Death Event Vs. Creatinine Phosphokinase") + xlab("Death Event") + ylab("Creatinine Phosphokinase")



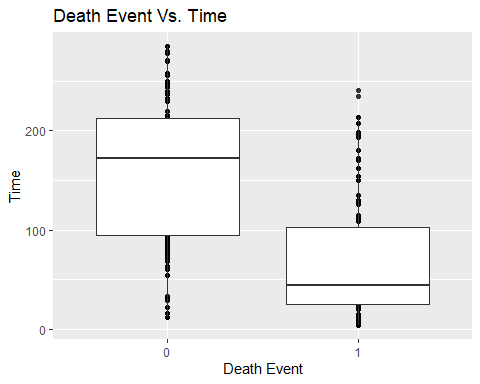
ggplot(heart\_factor\_df, aes(x=DEATH\_EVENT, y=ejection\_fraction)) + geom\_point() + geom\_boxplot() + ggtitle("Death Event Vs. Ejection Fraction") + xlab("Death Event") + ylab("Ejection Fraction")



ggplot(heart\_factor\_df, aes(x=DEATH\_EVENT, y=serum\_creatinine)) + geom\_point() + geom\_boxplot() + ggtitle("Death Event Vs. Serum Creatinine") + xlab("Death Event") + ylab("Serum Creatinine")



ggplot(heart\_factor\_df, aes(x=DEATH\_EVENT, y=time)) + geom\_point() + geom\_boxplot() + ggtitle("Death Event Vs. Time") + xlab("Death Event") + ylab("Time")



By looking at the mosaic plots, it appears that all the variables: anaemia, diabetes, high blood pressure, sex, and smoking did not play significant role in death by heart failure since the distribution is almost same percentage for each value.

By looking at the box plots, it appears that older people, people with low ejection fraction, people with high serum creatinine, and people with lower follow up time were prone to death by heart failure compared to the opposite groups. The level of Creatinine Phosphokinase did not contribute as a factor to death by heart failure.

I then calculated the correlation coefficient values of all variables with DEATH\_EVENT variable.

cor(heart\_df$DEATH\_EVENT, heart\_df$age)

## [1] 0.2537285

cor(heart\_df$DEATH\_EVENT, heart\_df$anaemia)

## [1] 0.0662701

cor(heart\_df$DEATH\_EVENT, heart\_df$creatinine\_phosphokinase)

## [1] 0.06272816

cor(heart\_df$DEATH\_EVENT, heart\_df$diabetes)

## [1] -0.001942883

cor(heart\_df$DEATH\_EVENT, heart\_df$ejection\_fraction)

## [1] -0.2686033

cor(heart\_df$DEATH\_EVENT, heart\_df$high\_blood\_pressure)

## [1] 0.07935106

cor(heart\_df$DEATH\_EVENT, heart\_df$serum\_creatinine)

## [1] 0.2942776

cor(heart\_df$DEATH\_EVENT, heart\_df$sex)

## [1] -0.004316376

cor(heart\_df$DEATH\_EVENT, heart\_df$smoking)

## [1] -0.01262315

cor(heart\_df$DEATH\_EVENT, heart\_df$time)

## [1] -0.5269638

Based on the correlation coefficients:

* age has low positive correlation with DEATH\_EVENT
* anaemia has no correlation with DEATH\_EVENT
* creatinine\_phosphokinase has no correlation with DEATH\_EVENT
* diabetes has no correlation with DEATH\_EVENT
* ejection\_fraction has low negative correlation with DEATH\_EVENT
* high\_blood\_pressure has no correlation with DEATH\_EVENT
* serum\_creatinine has low positive correlation with DEATH\_EVENT
* sex has no correlation with DEATH\_EVENT
* smoking has no correlation with DEATH\_EVENT
* time has high negative correlation with DEATH\_EVENT

Since the correlation coefficient can be trusted only with linear relationships, I created a logistic regression model and identified the statistical significance of the independent variables using p-values by calling summary() function.

heart\_glm <- glm(DEATH\_EVENT ~ age + anaemia + creatinine\_phosphokinase + diabetes + ejection\_fraction + high\_blood\_pressure + serum\_creatinine + sex + smoking + time, data=heart\_df, family = binomial)  
summary(heart\_glm)

##   
## Call:  
## glm(formula = DEATH\_EVENT ~ age + anaemia + creatinine\_phosphokinase +   
## diabetes + ejection\_fraction + high\_blood\_pressure + serum\_creatinine +   
## sex + smoking + time, family = binomial, data = heart\_df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2541 -0.5864 -0.2415 0.4620 2.7996   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.5952256 1.1860435 0.502 0.61577   
## age 0.0481262 0.0154658 3.112 0.00186 \*\*   
## anaemia -0.0533200 0.3558506 -0.150 0.88089   
## creatinine\_phosphokinase 0.0001997 0.0001747 1.143 0.25314   
## diabetes 0.1980819 0.3460965 0.572 0.56710   
## ejection\_fraction -0.0776583 0.0160315 -4.844 1.27e-06 \*\*\*  
## high\_blood\_pressure -0.1131988 0.3557769 -0.318 0.75035   
## serum\_creatinine 0.7170142 0.1800540 3.982 6.83e-05 \*\*\*  
## sex -0.4505779 0.4046368 -1.114 0.26548   
## smoking -0.0559680 0.4060595 -0.138 0.89037   
## time -0.0207200 0.0029555 -7.011 2.37e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 375.35 on 298 degrees of freedom  
## Residual deviance: 222.99 on 288 degrees of freedom  
## AIC: 244.99  
##   
## Number of Fisher Scoring iterations: 5

Based on the logistic regression model, age, ejection\_fraction, serum\_creatinine, and time variables were statistically significant variables since their p-values were less than 0.05. These same four variables showed correlation with DEATH\_EVENT variable when the correlation coefficients were calculated.

The next step I took was to split the dataset into training dataset and testing dataset with 80% and 20% sizes respectively. I then created a confusion matrix and calculated the accuracy of the model.

data\_split <- sample(1:nrow(heart\_df), 0.8 \* nrow(heart\_df))  
train <- heart\_df[data\_split,]  
test <- heart\_df[-data\_split,]  
  
res <- predict(heart\_glm, test, type="response")  
res <- predict(heart\_glm, train, type="response")  
  
confmatrix <- table(Actual\_Value = train$DEATH\_EVENT, Predicted\_Value = res > 0.5)  
confmatrix

## Predicted\_Value  
## Actual\_Value FALSE TRUE  
## 0 147 13  
## 1 26 53

accuracy <- (confmatrix[[1,1]] + confmatrix[[2,2]]) / sum(confmatrix)  
accuracy

## [1] 0.8368201

The accuracy of the logistic regression model was 83.68%.

I also used K Nearest Neighbors algorithm to predict the outcome. Since the dataset had 299 records, I used the K value of 17 which is an odd number and is close to the square root of 299. I then created a confusion matrix and calculated the accuracy of the model.

target\_category <- heart\_df[data\_split,11]  
test\_category <- heart\_df[-data\_split,11]  
  
library(class)  
test\_pred <- knn(train,test,cl=target\_category,k=17)  
  
confmatrix <- table(test\_category,test\_pred)  
confmatrix

## test\_pred  
## test\_category 0 1  
## 0 39 4  
## 1 10 7

accuracy <- (confmatrix[[1,1]] + confmatrix[[2,2]]) / sum(confmatrix)  
accuracy

## [1] 0.7666667

The accuracy of the K Nearest Neighbors Algorithm is 76.67%.

## Summarize the interesting insights that your analysis provided.

When I looked at the mosaic plots and box plots, I noticed that variables: age, ejection\_fraction, serum\_creatinine, and time showed significant impact on the DEATH\_EVENT variable.

When I calculated the correlation coefficient of all variables with DEATH\_EVENT variable, the same four variables: age, ejection\_fraction, serum\_creatinine, and time showed some level of correlation with DEATH\_EVENT variable.

When I used summary() function for the logistic regression model, the same four variables: age, ejection\_fraction, serum\_creatinine, and time showed statistical significance through low p-values. It is interesting to see that the same four variables appeared in all three methods, which confirms that they really were the appropriate variables that statistically affected the DEATH\_EVENT variable.

When we use our own understanding, we always think of parameters such as high blood pressure, cholesterol level, diet, and exercise as contributing factors for deaths due to heart failure. It is surprising that high blood pressure did not show any effects in my analysis. The data collected for this analysis is the data of patients that have had an incident related to heart, and what factors could impact survival after the incident. In that case, it makes sense for doctors to use this model to elongate the life of patients that have had an heart incident by looking at the risk factors.

## Summarize the implications to the consumer (target audience) of your analysis.

While working on the project, I realized that the target audience for this project are doctors and the patients who have already had a heart related medical incident. For such patients, doctors can use these prediction models to determine how to elongate the lives of the patients or prevent their death by looking at the factors. Age of the patient is an important factor. Older people are at higher risk, so they should be given higher level of care. People with low ejection fraction are at higher risk, so the doctors should prescribe medicines to these patients to increase their ejection fraction. People with high serum creatinine are at higher risk also, so the doctors should suggest patients to follow directions to reduce the serum creatinine levels. Some ways to do that are to drink more water, limit protein intake, limit sodium intake, and increase fiber intake. The longer the follow up time, there are more chances of the patient to survive. By taking these precautions, the doctors can help patients survive for longer time.

## Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.

One of the limitations of my analysis was the lack of data available. The dataset included only 299 rows, but this was the most appropriate dataset for my analysis. It would be great if more data is collected and made available for research, so that better models with more accuracy can be created.

Another limitation of my analysis was that the available data was of patients who have already had a heart incident. The analysis would help doctors to elongate the life of such patients by looking at the factors. Several people who are at high risk of getting heart attack do not know that there are ways that can be used to prevent the heart attack. Collecting data of such people that have not had a heart incident yet, and conducting analysis on the data will not only benefit the people at risk, but also the overall community.