Project Report: Heart Failure Clinical Dataset

Juliet Baldwin, Evan Juncal, Jason Au, Kevin Huynh

Summary

The goal of this investigation is to find correlations between sex, age, smoking habits, whether or not they have diabetes or high blood pressure as well as other related health factors and the associated level of risk they have for cardiovascular disease. We hope to compile these data analysis findings and accurately predict the age at which one could possibly die from the development of cardiovascular disease. This could in turn be used as a reference to prevent the risk of suffering through an unexpected heart attack while saving many lives through the process. Throughout our statistical analysis, we were able to uncover many unexpected findings such as how having diabetes and high blood pressure can play such a major role in heart failure while smoking plays a very minor role in comparison to the other factors in our statistical analysis of the data set. We identified descriptive statistics for each column and made comparisons of values grouped by averages and the event of death, finding the mean values that seemed to most influence CVD-related deaths. We were also able to discover many other findings relating to how the level of creatinine phosphokinase, ejection fraction, platelets, serum creatinine, and serum sodium in the blood of each sample patient correlated with the death rate of CVD patients. We used these differing values to calculate t-values which helped us test/predict whether or not a patient would possibly die from CVD. We also ran a K-Nearest Neighbors algorithm to run a prediction on how many of the major factors we identified, such as serum sodium, age and diabetes can affect CVD death rates with around 90% accuracy.

Table of Contents

Description

Cardiovascular diseases (CVDs) are a leading cause of death worldwide. According to the CDC, “One person dies every 36 seconds in the United States from cardiovascular disease [and] about 655,000 Americans die from heart disease each year—that’s 1 in every 4 deaths” (*Heart Disease Facts* 2020). In the United States, CVDs accounted for 23.5% of the 2.8 million deaths in 2017. Fortunately, the likelihood of developing CVDs can be greatly reduced by addressing behavioral risk factors such as unhealthy diet and obesity, physical inactivity, and harmful use of alcohol and tobacco. Therefore, it is important that individuals who are at high risk to develop CVDs need early detection and make necessary lifestyle changes to prevent heart diseases. The American Heart Association writes, “High blood pressure is sometimes called the "silent killer" because it often lacks obvious symptoms” (*Cardiovascular diseases* 2019). This emphasizes the importance of testing for cardiovascular diseases early on, and being able to predict what factors put people most at risk in order to personalize recommendations on how they can stay healthy. This motivated us to investigate the relationship between certain risk factors and death rates related to CVDs, to see the possibility of designing a method to predict a heart failure of an individual based on their risk factors. Additionally, we can provide graphics to further accentuate how such risk factors might affect CVD death rates in an easily digestible format for viewers and readers of our report to educate and understand how to potentially lower these rates and save lives.

Raw Data

For our project, we chose to analyze a dataset pertaining to heart failure. The dataset contains people's information related to their age, sex, blood pressure, whether or not they smoke, have diabetes, anemia, or hypertension as well as other related factors such as levels of CPK enzymes, platelets, serum sodium, and serum creatinine in the blood.

The dataset consists of categorical data and numerical data. The categorical data is represented as a boolean. The numerical data features contain both discrete and continuous data types. The 'age' data feature is an example of discrete data. The continuous data features and the boolean data features along with their meanings are outlined below.

Boolean features:

- Sex: Gender of patient. Male = 1, Female = 0
- Diabetes: 0 = No, 1 = Yes
- Anaemia: 0 = No, 1 = Yes
- High Blood Pressure: 0 = No, 1 = Yes
- Smoking: 0 = No, 1 = Yes
- Death Event: 0 = No, 1 = Yes

Continuous features:

- Creatinine Phosphokinase: Level of creatinine phosphokinase enzyme in the blood
  - Measured in mcg/L
  - Data range: [23, … , 7861]
- Ejection Fraction: Percentage of blood leaving the heart
  - Measured as percentage
  - Data range: [14, … , 80]
- Platelets: Measure of the amount of platelets in the blood
  - Measured in kiloplatelets/mL
  - Data range: [25.01, … , 850.00]
- Serum Creatinine: Level of creatinine in the blood
  - Measured in mg/dL
  - Data range: [0.50, … , 9.40]
- Serum Sodium: Level of sodium in the blood
  - Measured in mEq/L
  - Data range: [114, … , 148]

There is one data feature included in the dataset that we purposefully left out of our analysis. The data feature labeled 'time' captures the time of the death event in days, which is why we decided to leave it out. That is, the time at which the patient died, which is generally not useful, especially for prediction, since one of our goals is to predict whether a patient lives or dies based on certain factors. For these reasons, the time should not be used as an input of the model. No one using the model will be able to provide a time of death as input if they are seeking to predict what might cause the death of the patient that has not yet occurred.

Statistical Analysis

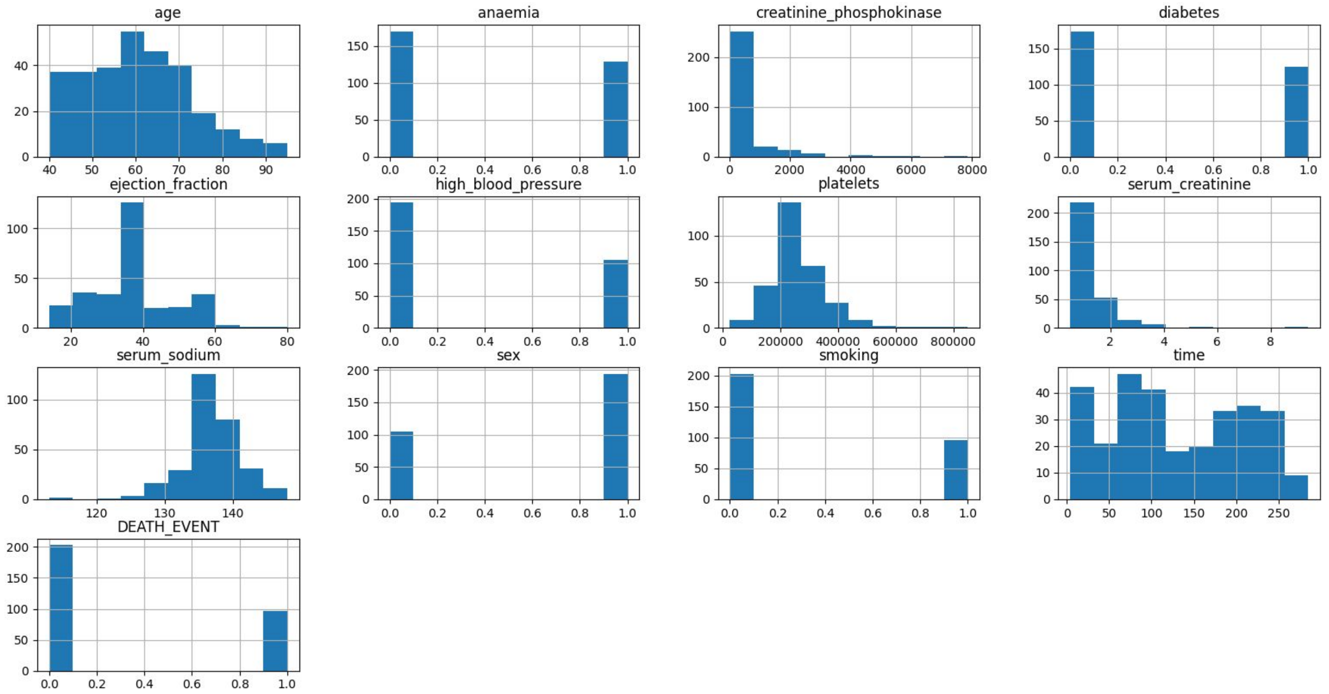


Figure 1: Histograms

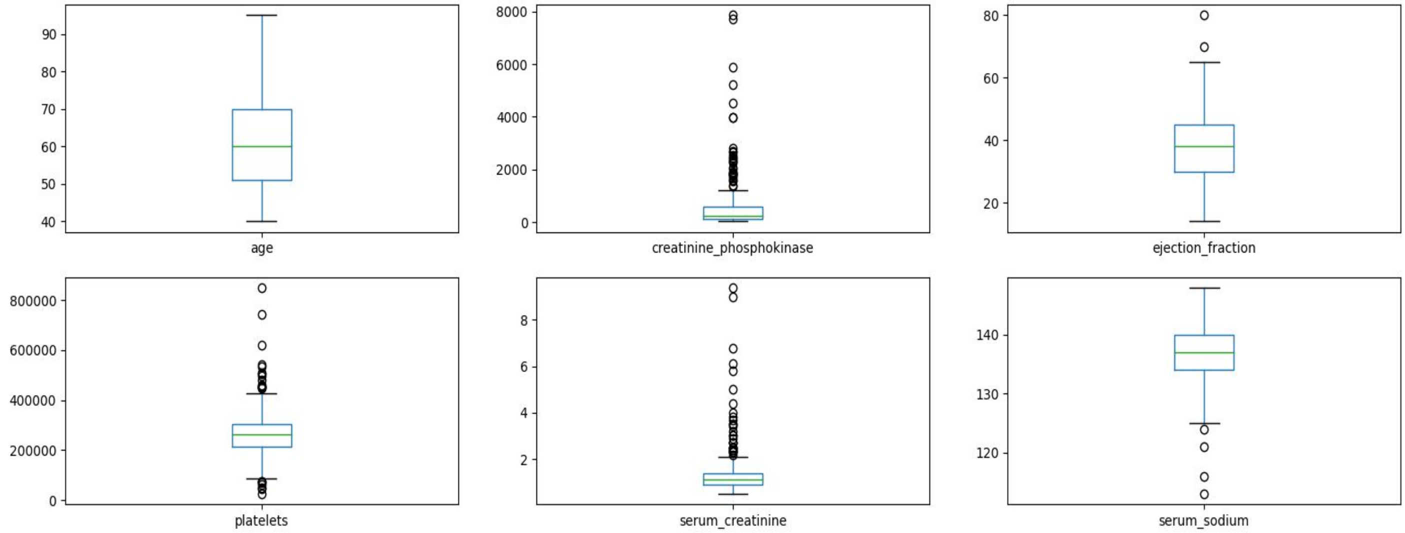


Figure 2: Box Plots

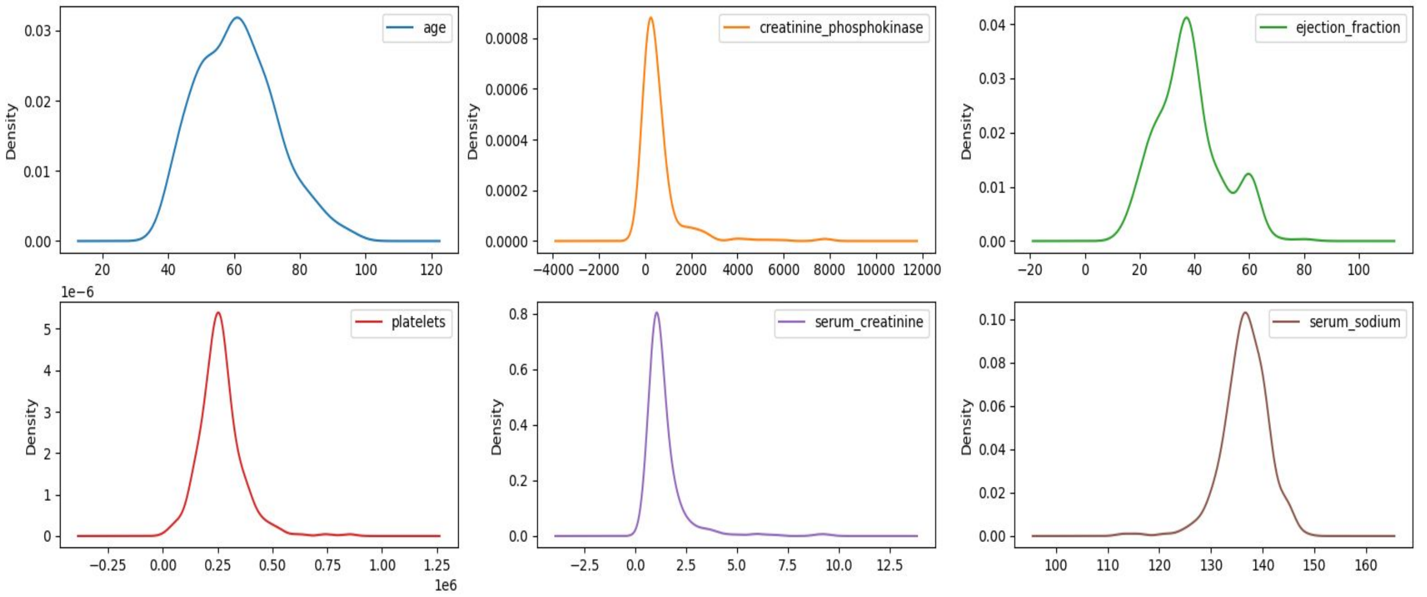


Figure 3: Density Plots

Based on the histograms and boxplots above, we are able to see that the age of people with cardiovascular disease tends to center around the ages of 40 to 70 while the much older aged people, around the ages of 75 to 90, would least likely suffer from cardiovascular disease. Additionally, we had more samples of men who had cardiovascular disease having almost about twice the sample size of women. Ironically, we had more samples of people who were not anaemic or diabetic, nor had high blood pressure or smoking habits but suffered from cardiovascular disease as opposed to the samples of people who did have those conditions and also suffered from cardiovascular disease.

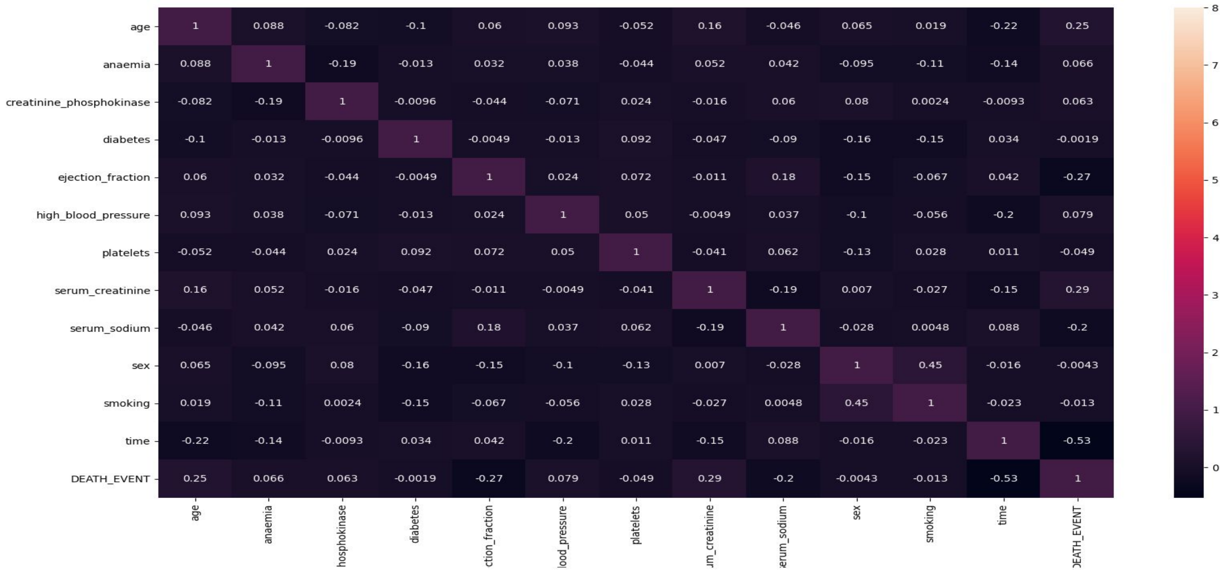


Figure 4: Correlation Heatmap of Data

Death Event	Age	Anaemia	Creatinine phosphokinase	Diabetes	Ejection fraction	High blood pressure	Platelets	Serum creatinine	Serum sodium	sex	smoking	time
0	58.8	0.409	540.054	0.419	40.266	0.325	266657	1.185	137.22	0.650	0.325	158
1	65.2	0.479	670.198	0.417	33.469	0.406	256381	1.836	135.38	0.645	0.313	71

Figure 5: Mean of values, grouped by whether the patient has died

From this, we can see that an older age has a higher risk of death on average, though not as big of an effect as one might think. Surprisingly, a higher amount of people on average had diabetes and survived compared to those who died, as well as ejection fraction (more blood leaving), platelets (which can cause blood clots in amounts that are too great, though having too little platelets can also cause issues), serum sodium (more sodium in the blood), and more smokers. Gender also had little effect, with women surviving only 0.005 more of the time on average. The rest of the data seems to follow as would be expected, as a greater percentage of those with anaemia (Decrease of red blood cells or hemoglobin), creatinine phosphokinase (Level of the CPK enzyme in the blood), high blood pressure (If a patient has hypertension), and serum creatinine (Level of creatinine in the blood) had a greater chance of death due to these factors.

```
              age         anaemia    creatinine_phosphokinase    diabetes  \
count   299.000000  299.000000            299.000000              299.000000
median   60.000000    0.000000            250.000000                0.000000
mean     60.833893    0.431438            581.839465                0.418060
std      11.894809    0.496107            970.287881                0.494067
var     141.486482    0.246122         941458.571457                0.244102
min      40.000000    0.000000             23.000000                0.000000
25%      51.000000    0.000000            116.500000                0.000000
50%      60.000000    0.000000            250.000000                0.000000
75%      70.000000    1.000000            582.000000                1.000000
max      95.000000    1.000000           7861.000000                1.000000

        ejection_fraction    high_blood_pressure      platelets  \
count       299.000000           299.000000            299.000000
median       38.000000             0.000000         262000.000000
mean         38.083612             0.351171         263358.029264
std          11.834841             0.478136          97804.236869
```

var	140.063455	0.228614	9565668749.448881
min	14.000000	0.000000	25100.000000
25%	30.000000	0.000000	212500.000000
50%	38.000000	0.000000	262000.000000
75%	45.000000	1.000000	303500.000000
max	80.000000	1.000000	850000.000000

	serum_creatinine	**serum_sodium**	**sex**	**smoking**	**DEATH_EVENT**
count	299.00000	299.000000	299.000000	299.00000	299.00000
median	1.10000	137.000000	1.000000	0.00000	0.000000
mean	1.39388	136.625418	0.648829	0.32107	0.32107
std	1.03451	4.412477	0.478136	0.46767	0.46767
var	1.070211	19.469956	0.228614	0.21872	0.21872
min	0.50000	113.000000	0.000000	0.00000	0.00000
25%	0.90000	134.000000	0.000000	0.00000	0.00000
50%	1.10000	137.000000	1.000000	0.00000	0.00000
75%	1.40000	140.000000	1.000000	1.00000	1.00000
max	9.40000	148.000000	1.000000	1.00000	1.00000

Figure 6: Descriptive statistics for each column

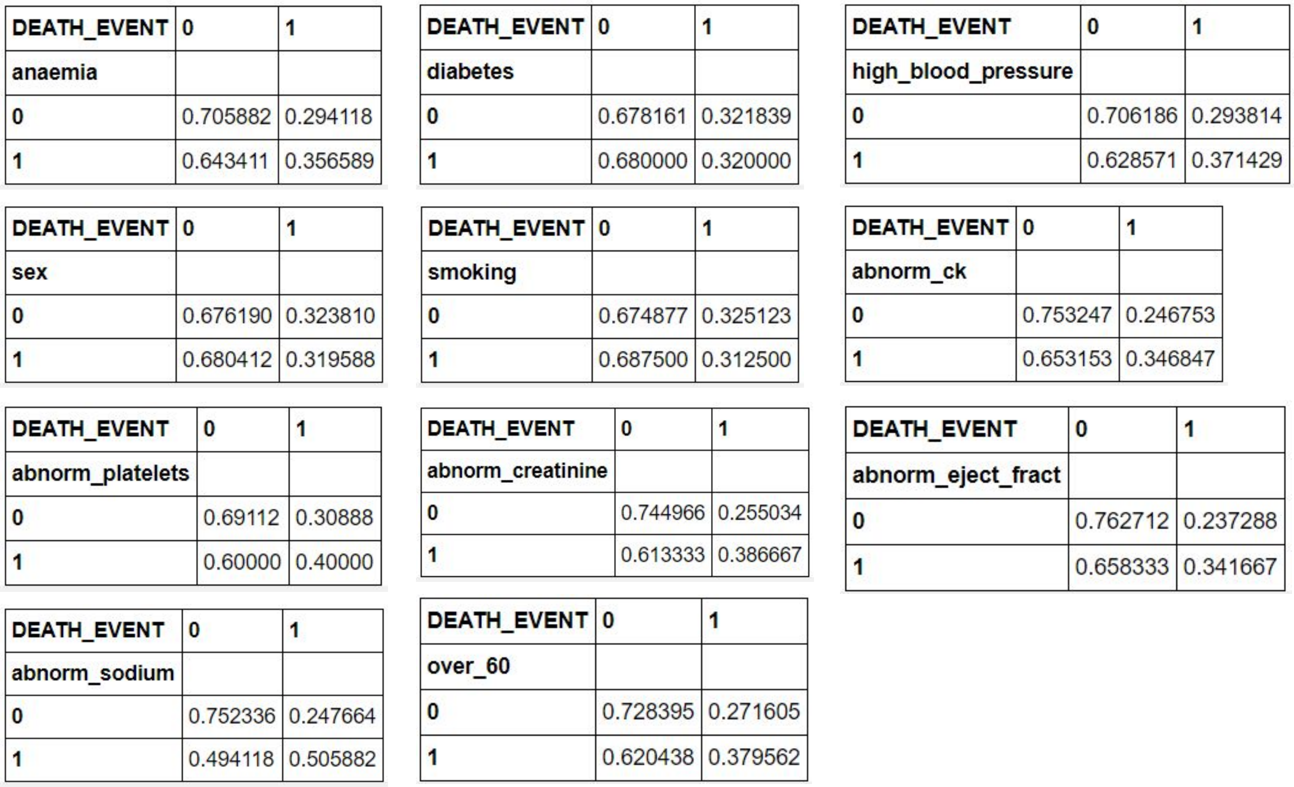
Outliers of creatinine_phosphokinase: 38 (col 1), 62 (col 52), 25 (col 60), 35 (col 72), 30 (col 103), 35 (col 134), 40 (col 171)
Outliers of ejection_fraction: 80 (col 64)
Outliers of platelets: 30 (col 105), 35 (col 109), 60 (col 296)
Outliers of serum_creatinine: 35 (col 9), 38 (col 28), 62 (col 52), 45 (col 131), 70 (col 217), 25 (col 228)

Figure 7: Outliers

```
CI: [63, 68] (mean age death event)
CI: [64, 70] (mean age of male death event)
CI: [58, 66] (mean age of female death event)
CI: [54, 66] (mean age with diabetes death event)
CI: [59, 76] (mean age of people who smoke death event)
CI: [55, 74] (mean age of people with high blood pressure death
              event)
CI: [62, 79] (mean age of people who smoke with high blood
              pressure death event)
CI: [55, 75] (mean age of people with diabetes and high blood
              pressure death event)
CI: [54, 76] (mean age of people with diabetes and smoking
death
```

```
                         event)
```

Figure 8: Confidence intervals



DEATH_EVENT	0	1
anaemia		
0	0.705882	0.294118
1	0.643411	0.356589

DEATH_EVENT	0	1
diabetes		
0	0.678161	0.321839
1	0.680000	0.320000

DEATH_EVENT	0	1
high_blood_pressure		
0	0.706186	0.293814
1	0.628571	0.371429

DEATH_EVENT	0	1
sex		
0	0.676190	0.323810
1	0.680412	0.319588

DEATH_EVENT	0	1
smoking		
0	0.674877	0.325123
1	0.687500	0.312500

DEATH_EVENT	0	1
abnorm_ck		
0	0.753247	0.246753
1	0.653153	0.346847

DEATH_EVENT	0	1
abnorm_platelets		
0	0.69112	0.30888
1	0.60000	0.40000

DEATH_EVENT	0	1
abnorm_creatinine		
0	0.744966	0.255034
1	0.613333	0.386667

DEATH_EVENT	0	1
abnorm_eject_fract		
0	0.762712	0.237288
1	0.658333	0.341667

DEATH_EVENT	0	1
abnorm_sodium		
0	0.752336	0.247664
1	0.494118	0.505882

DEATH_EVENT	0	1
over_60		
0	0.728395	0.271605
1	0.620438	0.379562

Figure 9: Conditional Probability

A useful way to incorporate health risk factors into the likelihood of a death event is to assume that the death event is due to a risk factor. The above figure presents the conditional probabilities of each risk factor against the odds of a death event. The risk factors include abnormal lab test results, existing health conditions, whether a person is over 60 years of age, sex, and whether a person smokes.

Lab Test	Normal Range
Creatine Phosphokinase	10 to 120 micrograms per liter (mcg/L) (Chen, 2019)
Ejection Fraction	50% to 70% (Burkhoff et al., 2003)
Platelet	150,000 to 450,000 platelets per microliter of blood (Hanke et al., 2010)

Serum Creatinine	0.84 to 1.21 milligrams per deciliter (74.3 to 107 micromoles per liter) (Hosten, 1990)
Serum Sodium	135 and 145 milliequivalents per liter (mEq/L) (Gao et al., 2016)

The above table shows the normal ranges for the lab tests. If a person's lab test result is outside of the normal range, then the person has an abnormal test result. Each of the above probabilities is calculated as follow:

Denote A : probability of a death event; B : probability of a risk factor

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

$P(A^c|B^c)$, $P(A|B^c)$, and $P(A^c|B)$ were also calculated using the above equation. From the conditional probabilities calculated above, we can identify which risk factors are more likely to lead to a death event. An abnormal serum sodium level is the risk factor that has the highest death rate while smoking, sex, and diabetes have some of the lowest rates. This approach assumes that a death event is due to a single risk factor and does not take into consideration that the risk factors are not mutually exclusive as a person can have multiple existing conditions and/or multiple abnormal test results. Although this approach gives us some general insights into the dataset, our results are not fully valid.

Based on the confidence intervals above that represent the population's mean age at which it would fall under, the average age of the entire sample population at which people die from cardiovascular disease would be between 63 and 68 meaning that the general population of people who have cardiovascular disease is estimated to die between the ages of around 63 and 68. Females tend to die earlier than males with a mean age interval of between 58 and 66 while the male population would live slightly longer with a mean age interval between 64 to 70. People with high blood pressure would tend to approximately die between the ages of 55 to 74 while people with diabetes tend to die a bit younger between the ages of 54 and 66. Surprisingly, the people who had smoking habits tend to live slightly longer with an interval age of 59 to 76 as opposed to people who had high blood pressure or diabetes. Additionally, the people who had smoking habits and were diabetic had similar intervals to people who were just diabetic with the mean death age interval of 54 to 76. Another unexpected finding was how people who smoked and had high blood pressure had a lower mortality with a mean death age of 62 to 79 which is higher than the people with only diabetes. However, not surprisingly enough, the people who were diabetic and had high blood pressure had a mean death age of 55 to 75. Based on these findings, we can assume that smoking does not have a significant effect on one's mortality from cardiovascular disease compared to the other factors.

μ_0: Mean age of people who smoke, have diabetes, and have high blood pressure

$H_0: \mu > 50$

$H_1: \mu \ngtr 50$

$t_0 = 3.48$

$t_{.05,\ 4} = 2.132$

$t_0 > t_{.05,\ 4}$

H_0 is rejected

After examining the different confidence intervals as well as other factor analyses, we wanted to test whether the true mean age of death for the population of people who both smoked, had diabetes, and had high blood pressure would be higher than the age of 50, meaning that we wanted to see if people who smoked, had diabetes, and had high blood pressure would likely die above the age of 50 based on the estimated true mean age of death for the population. The best approach to this phenomenon would be to conduct a hypothesis testing where H_0 would represent the hypothesis that μ (the true mean age of death of the population who smoked, had diabetes, and high blood pressure) would be estimated to be above the age of 50, and H_1 would represent the hypothesis that μ is not above the age of 50. According to our calculations, we computed that t_0 is 3.48 which is greater than the t-value of 2.132 with a confidence level of 95 percent and 4 degrees of freedom meaning that we reject our hypothesis of the mean death age of people who smoke, have diabetes, and have blood pressure would be above the age of 50. The possible reason for our results would be because the sample size of people who smoke, had diabetes, and had high blood pressure was significantly small with only 5 samples leaving only 4 degrees of freedom to work with. Due to this complication, the sample size was not able to meet the criteria of the central limit theorem, so our calculations are not fully valid.

Another approach we decided to look at was to compare the means of the numerical data features based on the death event. We wanted to see if there was a significant difference in the mean from the sample of people who died versus the sample of people who did not die. To test for a significant difference of means between the two groups, we used a T-Test. We performed a t-test on the following data features: Age, Creatinine Phosphokinase, Ejection Fraction, Platelets, Serum Creatinine, Serum Sodium. Based on the results, from the T-Tests, we rejected the null hypothesis for age, ejection fraction, serum creatinine, and serum sodium. For those given data features, we are confident that the difference in means is statistically significant rather than due to random chance.

Common Values for All T-Tests

N_0: Number of people from sample who did not die

N_1: Number of people from sample who died

$N_0 = 203$

$N_1 = 96$

$DF = 299 - 2 = 297$

$\alpha = 0.01$

$t_{0.01,\ 297} = 2.339$

Age T-Test

μ_0: Mean age of patients who did not die

μ_1: Mean age of patients who died

H_0: $\mu_0 - \mu_1 = 0$

H_1: $\mu_0 - \mu_1 < 0$

$t_0 = -4.521$

$-t_{.01,\ 297} = -2.339$

$t_0 < -t_{.01,\ 297}$

H_0 is rejected

Creatinine Phosphokinase T-Test

μ_0: Mean creatinine phosphokinase level of patients who did not die

μ_1: Mean creatinine phosphokinase level of patients who died

H_0: $\mu_0 - \mu_1 = 0$

H_1: $\mu_0 - \mu_1 < 0$

$t_0 = -1.083$

$-t_{.01,\ 297} = -2.339$

$t_0 > -t_{.01,\ 297}$

H_0 is not rejected

Ejection Fraction T-Test

μ_0: Mean ejection fraction of patients who did not die

μ_1: Mean ejection fraction of patients who died

H_0: $\mu_0 - \mu_1 = 0$

H_1: $\mu_0 - \mu_1 > 0$

$t_0 = 4.806$

$t_{.01,\ 297} = 2.339$

$t_0 > t_{.01,\ 297}$

H_0 is rejected

Platelets T-Test

μ_0: Mean platelets level of patients who did not die

μ_1: Mean platelets level of patients who died

H_0: $\mu_0 - \mu_1 = 0$

H_1: $\mu_0 - \mu_1 > 0$

$t_0 = 0.848$

$t_{.01,\ 297} = 2.339$

$t_0 < t_{.01,\ 297}$

H_0 is not rejected

Serum Creatinine T-Test

μ_0: Mean serum creatinine level of patients who did not die

μ_1: Mean serum creatinine level of patients who died

H_0: $\mu_0 - \mu_1 = 0$

H_1: $\mu_0 - \mu_1 < 0$

$t_0 = -5.306$

$-t_{.01,\ 297} = -2.339$

$t_0 < -t_{.01,\ 297}$

H_0 is rejected

Serum Sodium T-Test

μ_0: Mean serum sodium level of patients who did not die

μ_1: Mean serum sodium level of patients who died

H_0: $\mu_0 - \mu_1 = 0$

H_1: $\mu_0 - \mu_1 > 0$

$t_0 = 3.430$

$t_{.01,\ 297} = 2.339$

$t_0 > t_{.01,\ 297}$

H_0 is rejected

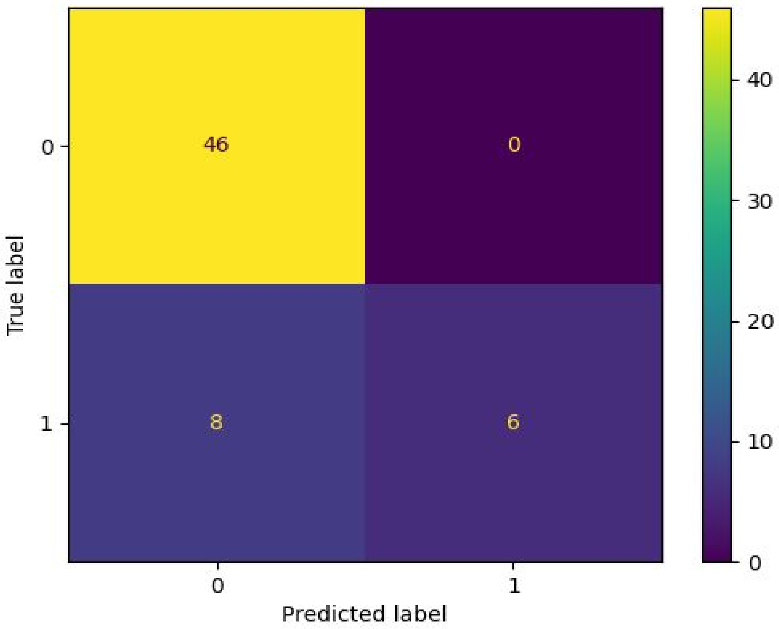


Figure 10: K Nearest Neighbors Top 5 Success Rates: [90.0, 90.0, 90.0, 88.33, 88.33]

K-Neighbors Prediction Results

We are comparing our predicted label (ejection fraction, serum creatinine/sodium, age, diabetes, hypertension, anaemia, sex, and smoking columns) to our true label (death event). From the above confusion matrix, we can see that we get 46 true negatives (TN), 0 false positives (FP), 8 false negatives (FN), and 6 true positives (TP), with an accuracy of around 90% for 5 runs. This means our Recall is R = TP/(TP + FN) = ~0.43 = 43%, which isn't the best. However, our Precision is calculated as P = TP/(TP + FP) = 1.00 = 100%, which is perfect precision. To compare the two, we can calculate the F-Measure (the harmonic mean between the two values), which is F = (2*R*P)/(R+P) = ~0.60 = 60%. Overall, we have a fairly good model for predicting what factors affect CVD death rates.

Implications

From our data, we can conclude that the factors which have the most effect on cardiovascular death rate appear to be age, serum sodium levels, and high blood pressure, as well as diabetes and gender to an extent. While the average for each gender was around the same age, the actual ranges of each gender’s death age vary somewhat significantly as seen above. Smoking, surprisingly, for this sample set, does not seem to have a great effect on the risk of cardiovascular death. We found that diabetes was one of the highest major factors that affect one's likelihood of early death due to cardiovascular disease. High blood pressure comes in at a close second for the most impact on people with cardiovascular disease leading to death. This leads us to believe that diabetes and high blood pressure are a top priority that many people should take into consideration when trying to prevent death from CVD or even developing CVD in general. We were also able to discover that there is a clear distinction for levels of serum sodium and serum creatinine in the blood as well as ejection fraction (percentage of blood leaving the heart) between the samples of people who died from CVD and people who did not die from CVD. Whereas, levels of creatinine phosphokinase and platelets in the blood were approximately equivalent between samples who died from CVD and samples who did not die from CVD according to our T-tests. This information can potentially be used by medical professionals to take blood tests of their patients and use it as a reference to possibly measure their level of risk for CVD or heart attacks. Additionally, we were able to calculate conditional probabilities for each of the different factors which is especially a powerful tool for medical professionals to accurately predict the likelihood of death.

Further Questions

A question brought up by our study that could be pursued more in depth in the future is whether other factors might have an impact on cardiovascular death rate, such as alcohol, amount of time people exercise, diet, genetics, etc. Another investigation we could have explored through our data analysis but simply did not have time to was how smoking, high blood pressure, and diabetes had an impact on levels of creatinine phosphokinase, ejection fraction, platelets, serum creatinine, and serum sodium in the blood of CVD patients. Furthermore, the same type of factors might be investigated in the future with a larger sample set in order to get better and more accurate results.

The results from an analysis on our sample seem to show smoking to some extent does not have an impact on cardiovascular disease. This could be because we had a much smaller sample size of people who smoked and had CVD as opposed to people who did not smoke and had CVD giving us a less accurate representation of how smoking can affect death due to heart disease. However, smoking still plays a major role in heart health. According to the American Heart Association, CVD accounts for about 800,000 U.S. deaths every year where nearly 20 percent of those deaths are due to cigarette smoking (Products, 2020). Therefore, cigarette smoking makes up only a fifth of the population of people who died from cardiovascular disease as opposed to other more major factors such as diabetes where at least 68 percent of people age 65 or older with diabetes have died from heart disease (Heart, 2015).

References

Ahmad T, Munir A, Bhatti SH, Aftab M, Ali Raza M. (2015, December). Survival analysis of heart failure patients: a case study dataset. Retrieved December 05, 2020, from https://plos.figshare.com/articles/Survival_analysis_of_heart_failure_patients_A_case_study/5227684/1

Burkhoff, D., Maurer, M. S., &amp; Packer, M. (2003, February 11). Heart Failure With a Normal Ejection Fraction. Retrieved December 06, 2020, from https://www.ahajournals.org/doi/full/10.1161/01.CIR.0000053947.82595.03

Cardiovascular Disease and Diabetes. (2015, August 30). Retrieved December 05, 2020, from https://www.heart.org/en/health-topics/diabetes/why-diabetes-matters/cardiovascular-disease--diabetes

Cardiovascular diseases affect nearly half of American adults, statistics show. (2019, January 31). Retrieved December 07, 2020, from https://www.heart.org/en/news/2019/01/31/cardiovascular-diseases-affect-nearly-half-of-american-adults-statistics-show

Chen, M. A. (2019). Creatine phosphokinase test: MedlinePlus Medical Encyclopedia. Retrieved December 05, 2020, from https://medlineplus.gov/ency/article/003503.htm

Gao, S., Cui, X., Wang, X., Burg, M., &amp; Dmitrieva, N. (2016, December 29). Cross-Sectional Positive Association of Serum Lipids and Blood Pressure With Serum Sodium Within the Normal Reference Range of 135–145 mmol/L. Retrieved December 06, 2020, from https://www.ahajournals.org/doi/full/10.1161/ATVBAHA.116.308413

Hanke, A.A., Roberg, K., Monaca, E. et al. (2010, May 18). Impact of platelet count on results obtained from multiple electrode platelet aggregometry (Multiplate™). Retrieved December 06, 2020, from https://doi.org/10.1186/2047-783X-15-5-214

Heart Disease Facts. (2020, September 08). Retrieved December 07, 2020, from https://www.cdc.gov/heartdisease/facts.htm

Hosten, A. O. (1990, January 01). BUN and Creatinine. Retrieved December 06, 2020, from https://www.ncbi.nlm.nih.gov/books/NBK305

Larxel. (2020, June 20). Heart Failure Prediction. Retrieved December 05, 2020, from https://www.kaggle.com/andrewmvd/heart-failure-clinical-data

Products, C. (2020, May 04). How Smoking Affects Heart Health. Retrieved December 05, 2020, from https://www.fda.gov/tobacco-products/health-information/how-smoking-affects-heart-health