Data Set

The data set used for this assignment was found online at kaggle.com and was originally published in the PLoS ONE journal titled: "Survival analysis of heart failure patients: A case study" [1]. The data set contains medical records of 299 patients experiencing heart failure collected at the Institute of Cardiology and Allied hospital Faisalabad in Pakistan. The records were collected from April to December 2015 and contains the following 13 attributes: **1 - age (integer),2 - anaemia (boolean), 3 - creatinine phosphokinase (integer), 4 - diabetes (boolean), 5 - ejection fraction (integer), 6 - high blood pressure (boolean), 7 - platelets (integer), 8 - serum creatinine (double), 9 - serum sodium (integer), 10 - sex (boolean), 11 - smoking (boolean), 12 - time (integer), 13 - DEATH_EVENT (boolean)**.

All non-boolean attributes were discarded before performing the association rule analysis on the data. Table 1 summarizes the attributes that were used and the interpretation of the boolean values. It is worth noting that for the target attribute of death event, a value of 0 can signify that the patient was alive at the end of the study or that the researchers were not able to locate the patient to continue the study.

Table 1: Processed heart failure data set

Column Number	Attribute	Value	
0	Anaemia	1 = anaemic	0 = not anaemic
1	Diabetes	1 = diabetic	0 = not diabetic
2	High Blood Pressure	1 = has hbp	0 = does not have hbp
3	Sex	1 = male	0 = female
4	Smoking	1 = smoker	0 = non-smoker
5	Death Event	1 = dead	0 = alive/censored

Problem Description

The goal of this assignment is to identify if there is a link between the ailments described in the table above and how strongly they are associated to the death of the patient due to heart failure.

Techniques Used

As previously stated, association rule analysis was used to identify associations in the data set. This technique can be used to identify relationships in transaction databases by using an apriori algorithm [2]. The rules produced by this analysis can then be ranked in terms of their "interestingness" based on three metrics: support, confidence, and lift [2]. Lift is a popular metric for measuring interestingness, however, standardized lift was used in this assignment due to the arbitrary scaling of lift within a max and min range which can differ for each rule. The equation for standardized lift is given by the following: $\mathcal{L}(A \Rightarrow B) = \frac{L(A \Rightarrow B) - \lambda}{v - \lambda}$ [2]. Standardized lift ranges from 0 – 1 and the higher it is for a given rule, the more interesting it is. The analysis was completed using the R programming language with the "arules" package [3] and the code for standardized lift was provided by Professor McNicholas.

Results

As mentioned in the problem description above, the goal of the assignment was to find associations between the ailments in Table 1, and if they resulted in the patient's death. Therefore, the right-hand-side of the association rule was fixed to true/yes (y5). After some trial and error, a support value of 0.0065, and confidence of 0.35 was used to generate 76 rules. The top 15 rules, sorted by standardized lift are shown below.

Table 2: Heart Failure Association Rules Results

	lhs		rhs	support	confidence	coverage	lift	count	slift
[1]	{n3,y1,y4}	=>	{y5}	0.006688963	1.0000000	0.006688963	3.1145833	2	1.000000000
[2]	{n3,y0,y4}	=>	{y5}	0.006688963	1.0000000	0.006688963	3.1145833	2	1.000000000
[3]	{n3,y4}	=>	{y5}	0.010033445	0.7500000	0.013377926	2.3359375	3	0.513736932
[4]	{n1,y0,y2,y4}	=>	{y5}	0.020066890	0.6000000	0.033444816	1.8687500	6	0.428571429
[5]	{n0,y1,y2,y4}	=>	{y5}	0.016722408	0.5555556	0.030100334	1.7303241	5	0.365079365
[6]	{y0,y2,y4}	=>	{y5}	0.020066890	0.5000000	0.040133779	1.5572917	6	0.285714286
[7]	{n3,y0,y1,y2}	=>	{y5}	0.020066890	0.5000000	0.040133779	1.5572917	6	0.285714286
[8]	{n1,n2,n4,y3}	=>	{y5}	0.060200669	0.4864865	0.123745819	1.5152027	18	0.266409266
[9]	{n3,y0,y2}	=>	{y5}	0.036789298	0.4782609	0.076923077	1.4895833	11	0.254658385
[10]	{n1,y2,y4}	=>	{y5}	0.030100334	0.4736842	0.063545151	1.4753289	9	0.248120301
[11]	{n1,y2,y3,y4}	=>	{y5}	0.026755853	0.4705882	0.056856187	1.4656863	8	0.243697479
[12]	{y2,y4}	=>	{y5}	0.046822742	0.4666667	0.100334448	1.4534722	14	0.238095238
[13]	{n3,y0,y1}	=>	{y5}	0.043478261	0.4642857	0.093645485	1.4460565	13	0.234693878
[14]	{n1,n4,y0,y3}	=>	{y5}	0.043478261	0.4642857	0.093645485	1.4460565	13	0.234693878
[15]	{y1,y2,y4}	=>	{y5}	0.016722408	0.4545455	0.036789298	1.4157197	5	0.220779221

From Table 2 it's evident that women who have diabetes and smoke, or women who smoke and are anemic are closely linked to passing away from heart failure. These two rules have a much higher standardized lift than the other combination of ailments. I've also highlighted the third rule which has a moderately high standardized lift which shows a strong relationship between women who smoke and death due to heart failure, but this rule can be categorized as a redundant rule because we can see from the first two rules that women who smoke are a subset of their itemset.

The other rules in Table 2 were not considered to be too interesting because of their significantly lower standardized lift, however, one thing that's worth noting is that almost all of the rules contain smokers on the antecedent side which can illustrate how much of an impact smoking has on the heart health in the Pakistani population where this data set was collected. One thing that I found odd from these results is that women were shown to have a higher link to death due to heart failure, despite the fact that research from Azad, Nahid et al [4] states otherwise. This is likely due to the low sample size and certain biases in the data set, such as exposure to second-hand smoke are not mentioned in the original research related to the data set [1].

References

[1] Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA (2017) Survival analysis of heart failure patients: A case study. PLoS ONE 12(7): e0181001. https://doi.org/10.1371/journal.pone.0181001

[2] Sharon McNicholas (2019). Association Rules - Lecture Notes. Hamilton, ON: McMaster University, Department of Mathematics and Statistics.

[3] Michael Hahsler, Christian Buchta, Bettina Gruen and Kurt Hornik (2018). arules: Mining Association Rules and Frequent Itemsets. R package version 1.6-2. https://CRAN.R-project.org/package=arules

[4] Azad, Nahid et al. (2011) *Gender differences in the etiology of heart failure: A systematic review.* Journal of geriatric cardiology : JGC vol. 8,1: 15-23. doi:10.3724/SP.J.1263.2011.00015