Assignment 1: Exploring the burn dataset

Michael Beauchamp

UMGC Data 630: Machine Learning

Fall 2021

Dr. Firdu Bati

**Introduction**

According to the World Health Organization, “an estimated 180,000 deaths every year are caused by burns” (Burns, 2021). The WHO describes a burn not only as being caused by a thermal burn such as flame or hot water, but radiation, radioactivity, electricity, friction or contact with chemicals (Burn, 2021).

The majority of burns that result in death are from lower to middle income countries and according to the WHO, “females have slightly higher rates of death from burns compared to males” (Burns, 2021). Children, similarly, are at a higher risk to suffer burn injuries which can lead to sever disfigurement, disability and extended hospitalizations (Burns, 2021). Logic would say that those statistics would be accurate as women, historically, tend to do more of the cooking in the home exposing them to more risk of burns. Similarly, children will be more prone to accidents involving fire because of their lack of knowledge and experience dealing with not only fire, but chemical burns as well.

This analysis is going to explore the possibility of burn deaths and severity related to gender and age to substantiate or refute the statements made by the WHO. Using the Apriori method is useful in this analysis because the resulting association rules create cases in which the probability of itemsets occur. Those rules can be used to determine that if a person dies from a burn wound, is there a correlation between their age or gender and we can use to make an assumption for future cases.

**Analysis**

The burn dataset being used for this analysis is stored on GitHub, however the origin of the data is not given. The dataset contains 9 variables as outlined in appendix figure 1 with 1,000 records. The 9 variables include:

1. Identification code
2. Burn Facility
3. Hospital Discharge Statue
4. Age at admission
5. Gender
6. Race
7. Total burn surface area
8. Burn involved inhalation injury
9. Flame involved in burn injury

The analysis will center on the variable hospital discharge status to determine if the gender, race, burn surface area or facility have any relationship to whether a patient is alive or deceased.

Reading in the dataset to a variable “burn”, the command str(burn) and summary(burn) highlight the descriptive and quantitative aspects of the attributes. In figure 2 under the summary command the only attributes with values other than 0 or 1 are facility, age and total burn surface area excluding the ID attribute. Under the age attribute, the youngest patient was .10 years old while the oldest was 89.70. Running barplot(table(burn$GENDER)) creates a bar plot showing that 300 records are variable 0 while 700 are variable 1. Cross referencing that with figure 1, we know that there are 300 female patients and 700 male patients.

Figure 4 is a histogram of the attribute Age is the burn dataset and tells us that being skewed right, most of the patients are on the younger side. Indeed, most of the patients are under 10 years old. The target variable, Death was plotted into a bar plot in figure 5 and shows that the vast majority, 850 patients, survived their wounds while 150 patients died of their wounds. Another interesting fact is that most of the patients were treated at facilities 1-10, as the chart in figure 6 shows that it is also skewed right.

The variable Total burn surface area is a useful variable as it is the only one, other than Death, that provides us with any indication as to the severity of the wounds. Looking at figure 7, the bar plot shows that the majority of the burn surface areas are small. There are some increases in the percentage of burn surface area periodically as you move right along the x axis of the table. For instance, there is a jump at 35% and another right around 90%.

To run the Apriori method on the burn dataset, we must convert the int attributes to factor datatype and delete those values, such as ID, that are not pertinent to the analysis. Figure 8 shows the code used to complete the data preprocessing and the results showing the code converted the attributes correctly. In this dataset, most of the columns were converted to straight factor using burn$columnName<-factor(burn$columnName) code while age and total burn surface area were converted using discretize command to put records into 6 bins.

Running the Apriori method with the default parameters of support at 10% and confidence at 80% we return 175 rules as seen in figure 9. The minimum and maximum length are 1 and 10 respectively. Inspecting the first 10 rules, some of the support and confidence percentages are low however, the rules have good lift values indicating that the rule is acceptable. For example, rule 3 says that when there was an inhale injury there was a flame present. The support value is 11.6%, confidence at 95.0% and a lift value of 1.79 all indicate that there is correlation between the two attributes in the itemset. However, to see what rules were created that involve the target variable we need to modify the parameters of the method.

Increasing the support value to .4 and the confidence value to .9 results in 28 rules. Figure 10 shows that code and the output of the rules that were created with these specific parameters. Rule 3 tells us that 72% of the records had a total burn surface area between .1 and 16.4%. With a confidence of 95%, we know that the majority of the those records also had a Death value of 0 indicating they survived. With a lift value of 1.12 we can say that patients are likely to survive given their burn surface area is between .1 and 16.4%. Cleaning up the rules by sorting and pruning redundant rules we find that there are 5 rules remaining. The code for sorting and cleaning is in figure 11. Figure 12 shows those rules plotted in a scatterplot. The scatterplot shows that all the rules have a lift value over 1, high confidence and support levels. The graph is effective in visualizing the remaining rules and providing a quick reference to their results.

**Results**

To determine if any rules can be created about deceased patients, we target the variable Death = 1. If we run the code with this variable on the right-hand side of the method, we do not return any rules even with the parameters set very low, see figure 13. We can return 2 rules with the same parameters however if we switch the target variable to the left-hand side of the method. In figure 14, we show the code and the output of switching the target variable to the right-hand side of the method. Rule 1 is the only one with support, confidence and lift values acceptable to be used. The target variable, Death=1 is in 13% of the records while 86% of those records also have Flame=1 which means there was a fire present in those case with a deceased patient. The rule has a lift value of 1.6 so we know that it is likely that when the patient is decease, there was a fire present.

The dispute the claims made by the WHO, we needed to look at the rules where Death=1. With the lack of rules that were created using the Apriori method, we cannot in fact contradict their claims. However, we can explore how accurate the WHO claims are by pursuing the rules where Death=0 indicating that the patient survived their burns. With Death=0 on the right-hand side of the method, support at .1, confidence at .8 and a minimum length of 2 there were 137 rules. Sorting these rules by the lift the first 20 rules all had a confidence level of 1 which means all the conditions on the left-hand side of the method all occurred when the variable Death=0. For example, in figure 15, rule 20 has a length of 4 with Age [.1,15], total burn surface area [.1-16.4] and inhale injury equal to no. Those variables combined for 24% of total transactions of which 100% had patients who survived their wounds. We know that with a lift value of 1.17 the rule is acceptable. This rule helps to confirm the claim of the WHO that children are susceptible to burn injuries.

Another claim by the WHO is females are higher risk from death by burns. Even though the method could not show any rules where the patient died, we can try to explore gender and burn relations. Figure 16 shows the Apriori method run with patients who survived and gender equal to female. The method with support at .1 and confidence at only .6 returned 1 rule. This rule is not acceptable because the lift value was only .98. Since only 1 rule was created, we can look back at figure 15 to see that gender=0 also occurs where Age [.1-16.4] and inhale injury=0 with support of 10%, confidence of 1 and a lift 1.17 which makes the rule acceptable.

Exploring another interesting value, Flame=0, can tell us how many of the patients were injured by burns that were not caused by a flame. The Apriori method created 7 acceptable rules with Flame being listed on the right-hand side of the method. These rules support the claims from the WHO that burns do not have to be caused by a flame. For example, rule 1 in figure 17 says that Age [.1-15] and total burn surface area [.1-16.4] make up 11% of all records. Of those records, 87% of them do not involve flames. With a lift value of 1.85 the rule is acceptable.

**Conclusion**

The WHO states that females and children and more susceptible to burns than the rest of the population in middle to low-income environments. Our analysis shows that there is no relationship between deaths and the rest of the variables in the dataset and we are not able to prove or disprove that females or children die due to burns at any rate different from males or adults. We can draw conclusions that patients who survived their burns were more likely children and male in contrast to the stated objective.

There are several limitations to the dataset that hindered the analysis. First, the dataset could have been larger, with only 1,000 records the data seemed to be skewed toward male patients. With a larger set, we may have been able to gather more association rules around gender. The same could be said about the number of survivors to the number of deceased. A timeline would also have been beneficial to the analysis. We could have looked at how the burns trend over time by gender, age or facility. This would have allowed us to draw conclusions as to whether or not the rate of burn victims is increasing or decreasing.

The dataset did not include any documentation as to where it was gathered from. Without knowing this piece of information, we cannot connect it to the WHO statements as they primarily speak on lower to middle economic areas. If the dataset included any sort of income or class status, we could have been able to segregate the patients based on this distinction. If we were to have the locations of the facilities, we could have at least been able to assume the locations that the patients as logic says they wouldn’t travel far to seek help. We then could have used those locations to gather general economic data on those citizens close to the facilities.

The WHO also spoke a lot about disabilities caused by the severity of burns. From our dataset, we can only speculate on the long-lasting effects of the damage occurred due to the burn from the total burn surface area variable. The patient data could have used an attribute related to severity of the wounds, such as the degree, that would have helped to also corroborate the claims made by the WHO on disabilities and lasting injuries. Race is another example of a variable that could have been more descriptive which would provide more analysis as to whether burns affect different races.

Length of time that the patient was in the hospital would also be a effective attribute for creating association rules. We could analyze the relationship between the total burn surface area and the length of stay to determine severity. If we also had a variable on severity, we could cross the surface are and severity to estimate the length of stay and likewise, whether the patient lives or dies due to their injuries.

The Apriori algorithm functions best on large data sets. The burn dataset is only 1,000 records over a timeframe that we do not know. If we could gather more records for burn patients, we may be able to obtain more rules on the relationship between patients that died due to their wounds and the rest of the variables. Once we have the association rules tied to the large dataset, we could make recommendations as to which part of the population is the more likely to suffer a burn and how server that burn could be.

**References**

Burns. (n.d.). www.who.int. Retrieved September 5, 2021, from https://www.who.int/news-room/factsheets/detail/burns#:~:text=Key%20facts%20An%20estimated%20180%20000%20deaths%20every.

**Appendix**

Table

Description automatically generated

Figure 1: Detailed description of the variables used in the burn dataset

Text

Description automatically generated with medium confidence

Figure 2: str and summary commands run of the burn dataset

Chart, shape

Description automatically generated

Figure 3: Histogram showing that there are 300 female patients and 700 male patients

Chart, histogram

Description automatically generated

Figure 4: Histogram showing that the age variable is skewed right, most patients are < 10 years old

Chart, shape, square

Description automatically generated

Figure 5: Histogram showing that 850 people survived their wounds while 150 died of their wounds

Chart, histogram

Description automatically generated

Figure 6: Histogram showing a skew right on the facility data

Chart, histogram

Description automatically generated

Figure 7: Frequency of burn area by %

Text

Description automatically generated

Figure 8: Data pre-processing, delete the ID column and change datatypes to factor

Table

Description automatically generated with medium confidence

Figure 9: Apriori method run on the burn variable results in 175 rules

A picture containing text

Description automatically generatedFigure 10: Apriori method run on the burn dataset with support at .4 and confidence at .9

Text

Description automatically generated

Figure 11: Sort an pruning redundant rules

Table

Description automatically generated

Figure 12: Scatterplot of the 5 rules resulting from pruning

Graphical user interface, text, email

Description automatically generatedFigure 13: Death=1 on the rhs of the method results in no rules even with parameters set low

Text

Description automatically generatedFigure 14: Death=1 on the lhs of the method results in 2 rules

Table

Description automatically generated

Figure 15: Apriori method run to find rules where death=0 sorted by the lift value

Graphical user interface, text, email

Description automatically generated

Figure 16: Apriori method run to find rules with death=0 and gender=0

Text

Description automatically generated

Figure 17: Apriori method run to find rules with flame=0