**Aluno: Keslley Lima da Silva**

**Matrícula: 00312107**

**Estudo Dirigido 01 - Mineração de dados usando regras de associação**

**Q1: Informe o número de registros do arquivo, o número de atributos, e o tipo de dado de cada atributo?**

O dataset possui 2201 registros, 4 atributos (Class, Sex, Age, Survived), no qual Class é do tipo de dado com 4 opções (Crew, 1st, 2nd, 3rd), Sex 2 opções (female, male), Age 2 opções (Adult, Child) e Survived 2 opções (yes, no).

**Q2: Diga pelo menos dois fatos que o summary revelou sobre os passageiros do Titanic**

Por meio do comando é possível concluir que grande parte das pessoas dentro do navio morreram, um total de 1490 que representa 67,7% do total de pessoas presente no navio.

A maioria das pessoas eram homens e adultos, cerca de 95% são adultos e os outros 5% são crianças. Já em relação ao sexo, 78% eram homens.

O navio tinha uma equipe de tribulação muito grande quando comparado ao número de passageiros. Cerca de 40% das pessoas que estavam no navio eram da tripulação.

O menor número de passageiros se encontravam em segunda classe (285 pessoas) e não na primeira (325 pessoas).

**Q3: Quantas regras o algoritmo apriori encontrou para este conjunto de dados?Analise a saída, e veja qual o valor de suporte e de confiança mínimos adotados pelo algoritmo apriori(parâmetros default)**

Foram encontrado 27 regras.

O valor de suporte é 0.1 (10%).

O valor de confiança é 0.8 (80%).

**Q4: Observe as duas primeiras regras listadas após o ordenamento.Você considera que essas regras são uma associação relevante? Justifique sua resposta.**

Não são relevantes pois são muito óbvias, isso é devido ao fato de ter um suporte muito alto e nesse caso essa duas regras acabam representando apenas distribuição do dataset, que é sobre o fato da maioria das pessoas serem homens e adultos. Ou seja, de forma geral essas duas primeiras regras não contribuem em nada para responder a pergunta de negócio.

**Q5: Expresse em português como você lê as regras [2] e [3], incluindo a interpretação do suporte e da confiança.**

regra [2]: Se o sexo da pessoa é homem então é um adulto.

O suporte diz que a frequência de registros que tem o atributo homem e adulto no dataset é de 75%.

A confiança diz 95% dos adultos também são homens.

regra [3]: Se a pessoa morreu então é um adulto.

O suporte diz que a frequência de registros que tem o atributo sobreviveu sendo não e sexo sendo adulto no dataset é de 65%

A confiança diz que 96% dos adultos também são pessoas que não sobreviveram.

**Q6: Análise as demais regras, e selecione uma que você considera relevante em relação ao problema de negócio definido. Informe o número da regra escolhida, e justifique sua escolha considerando as métricas de suporte e confiança.**

Regra [20], que é se passageiros da 3º classe e sexo masculino então não sobreviveram. Os principais motivos da escolha dessa regra é:

* Possui um suporte baixo (19%) e uma confiança relativamente alta (82%), o que idealmente é mais interessante.
* O fato de um baixo suporte demonstra que o resultado da confiança e até mesmo a conclusão da regra não é óbvio como regra [2], se homem então adulto, pois possui uma frequência de registro na base baixa, sendo menos de ⅕ da base que possuem os atributos 3rd e sexo masculino.
* Em relação a confiança é possível afirmar que 82% dos homens da 3º classe não sobreviveram. O que é um informação importante para responder quais fatores levaram ou não a morte.
* No caso baseado nessa regra é possível identificar que fatores que levaram morte dos passageiros é o fato de serem da 3º classe, que é um espaço geralmente ocupado por pessoas com situação econômica menos favoráveis e isso pode levar ação de analisar estrutura do navio para saber se a região que estava 3º classe não era mais longe e/ou difícil acesso dos botes, coletes salva vidas e saídas de emergência.
* Outro fator é em relação ao sexo ser masculino, ao meu ver não é um fator tão interessante devido ao fato de 80% dos passageiros serem do sexo masculino, porém é esperado devido ao maior número de homens que a probabilidade de morrer mais homens do que mulheres é maior, visto que tem mais homens do que mulheres.

**Q7: Observe as quatro primeiras regras listadas após o ordenamento por confiança. Você considera que, porque estas regras têm confiança máxima, elas são mais relevantes para sua questão de negócio? Justifique sua resposta.**

Apenas analisando o valor de confiança de forma isolada não considero relevantes para responder questão de negócio, pois caso possua um suporte alto essas regras mesmo com uma confiança máxima representa associações óbvias. Esse cenário é exatamente o que acontece nas 4 primeiras regras listas, que apesar de uma confiança alta, são óbvias, porque estão representado relações que acontecem devido ao grande número de registros com esses atributos, como exemplo se homem então adulto e a regra se não sobreviveu então adulto. Essas duas regras provavelmente são devido ao fato de 67% dos passageiros não sobreviverem e ao fato de 80% eram do sexo masculino.

**Q8: Para suporte e confiança iguais a 90%, alguma regra relevante foi**

**retornada?**

Não. A regra gerada não contribuem em nada para responder a questão de negócio.

**Q9: Baixando o suporte e confiança mínimos, apareceram mais regras?**

Sim, aparecem agora 62 regras.

**Q10: Elas são relevantes para sua questão de negócio? Explique sua resposta.**

As regras de [1] a [5] não são relevantes para responder questão de negócio, pois ~são muito óbvias. Isso é devido ao fato da listas está ordenado em ordem decrescente pelo valor do suporte.

**Q11: As regras resultantes parecem ser orientadas ao problema de negócio? Explique sua resposta.**

Sim, todas as 7 regras geradas são orientados ao problema do negócio. Porém em um cenário perfeito nesse conjunto de regras conteria também o lado direito da regra sendo sobreviveu = sim, pois questão de negócio não é apenas os fatores que levaram a não sobreviver e sim fatores que levaram a sobreviver ou a morte. Portanto, tendo do lado direito da regra sim e não para atributo sobreviveu seria mais abrangente para responder ao problema.

**Q12: Por que não temos regras que explicam a sobrevivência? Que parâmetro você poderia alterar para conseguir regras que explicam a sobrevivência?**

Acredito que não aparecem regras que tem o lado direito sendo sobreviveu = sim devido ao suporte, abaixando apenas o suporte deve aparecer regras com o lado direito sendo sobreviveu = sim. No caso abaixando o suporte para 0.15 (15%), antes estava 30%, já aparece uma regra que é se sexo = feminino então sobreviveu. Abaixando o suporte para 10% aparece mais outra regra que é se sexo = feminino e idade = adulto então sobreviveu

**Q13: Existem regras que lhe parecem redundantes entre si? Quais?**

As regras [3] e [7] parecem redundantes, pois as regras [1] e [2] já resulta na mesma relação que a [3]. O mesmo vale para as regras [4], [5] e [6] em relação a regra [7]

**Q14: Comparadas com as regras da execução anterior (Q11), cite dois exemplos distintos de redundâncias que foram eliminadas? Explique sua resposta em termo dos itens que aparecem nas regras, e suas métricas.**

As regra se sexo = masculino e idade = adulto então sobreviveu = não e Se classe = tripulação e sexo = masculino e idade = adulto então sobreviveu = não. Em relação a primeira ela tinha suporte de 60% e confiança de 80%. Já a segunda suporte 30% e confiança 77%. Acredito que em ambos casos as regras foram removidas devido ao valor de confiança e suporte muito próximos e também devido ao fato de utilizar os mesmo atributos de outras regras.

**Q15: Compare as regras [4] e [17]. Observe que têm confianças parecidas, mas lifts distintos. O que estas regras nos dizem sobre a sobrevivência?**

A regra [4] diz que a probabilidade dos itens co-ocorrerem é alta, tendo então uma alta relação de causa-consequência entre sexo = feminino, idade = adulto e sobreviveu = sim. Portanto, regra 4 é ótima para responder questão de negócio, pois tenho baixo suporte, alta confiança e lift.

Já a regra [17] com lift próximo de 1 diz que a ocorrência dos itens são independentes, ou seja, que não existe uma relação de causa consequência (se-então) entre idade = adulto e sobreviveu = não. No caso deu uma confiança alta, semelhante a da regra [4], devido ao alto suporte.

**Q16: Compare as regras [3] e [18]. Embora elas tenham confiança um pouco distintas, o que os respectivos lifts nos dizem sobre a não sobrevivência?**

A regra [3] diz que a probabilidade dos itens co-ocorrerem é alta, tendo então uma alta relação de causa-consequência entre classe = 2nd, sexo feminino e sobreviveu = sim. Portanto, essa regra diz, devido valor do lift, que existe uma real relação entre os da esquerda com da direita e então ótima para ser usada para responder questão de negócio.

Já a regra [18], classe = 2nd e sobreviveu = não, diz que a presença, o fato de ser classe= 2nd, diminui a probabilidade de ocorrência de sobreviveu = não. Logo pode ser uma regra interessante de um fator que influencia na sobrevivência.

**Q17: Com base nestas regras, liste 5 coisas que podem ser inferidas dos dados sobre sobreviventes e não sobreviventes do Titanic com o uso de regras de associação.**

O fato de ser da 2º classe é um fator que contribuiu para sobrevivência independente da idade e sexo.

Todas crianças que estavam na 2º classe sobreviveram.

Mulheres ricas, que estavam na primeira e segunda classe, possuem alta chance de sobreviverem.

Homens que estavam na 2º e 3ª classe possuem alta chance de morrerem.

Pessoas da tripulação tem alta chance de morrerem.

**Q18: crianças da segunda classe tiveram mais chances de sobrevivência do que as da primeira classe?**

Não se pode chegar essa conclusão pois não foi gerado nenhuma regra que relaciona criança, 1º classe e o fato de sobreviveu ou não. Um dos motivos pode ser devido ao suporte ser 0.01 e as regras que fazem essa relação não atender ao suporte mínimo. Também pode ser devido ao valor da confiança mínima que não foi atingido.

**Q19: Responda novamente à questão Q18 utilizando as métricas de suporte/confiança para justificar sua resposta.**

Todas crianças da segunda e primeira classe sobreviveram, pois possuem confiança de 100%. Além disso, mostra alto lift, o que reforça que essas duas são regras são interessantes para responder a questão de negócio.

**Q20: E as crianças da terceira classe? Como sua situação se compara (a) à das demais crianças? (b) dos adultos da primeira classe?**

1. As crianças da terceira classe tem uma confiança de 65% para a regra de classe = 3rd, idade = criança então sobreviveu = não. Sendo assim, grande parte das crianças da 3º classe não sobreviveram, totalmente diferente do que aconteceu com as crianças da primeira e segunda classe.
2. Em relação aos adultos da terceira classe a regra possui 75% de confiança, ou seja, grande possibilidade de adultos da terceira classe não sobreviverem.

**Q20: Qual a regra com suporte mais alto? Qual suporte resultaria em regras com maior confiança?**

A regra [17] possui o maior suporte entre as regras, que é de 0.65. Um suporte de máximo de 10% pegaria as 5 regras com maior confiança.