**Aluno: Keslley Lima da Silva**

**Matrícula: 00312107**

**Estudo dirigido sobre classificação no weka (opcional)**

### DATASET TITANIC - RESPOSTAS

**Q1:**

A 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).

**Q2:**

Acredito que a classe na qual a pessoa está é o principal fator que determina sobrevivência, pois pela visualização em gráfico maioria da 1 classe sobreviveu e já a 3 classe e tribulação grande parte morreu.

**Q3:**

Acredito que existe uma desbalanceamento na proporção de registros, pois na base de dados o número de registros de pessoas que não sobreviveram é 2 vezes maior do que as que sobreviveram. Logo o modelo preditivo pode ter dificuldade (menor acurácia) em classificar quando dado certas características a pessoa vai morrer.

**Q4:**

Pelo algoritmo zeroR ele vai classificar todos novos registros como não sobreviventes, devido maior número de registros no dataset de treinamento ser não sobreviventes.

**Q5:**

Sim, o resultado foi “ZeroR predicts class value: No”

**Q6:**

A acurácia foi melhor em relação ao ZeroR, pois no OneR obteve 77,6%.

**Q7:**

A precisão aumentou para 78% e o recall caiu para 91%.

**Q8:**

A precisão 73% e o recall caiu para 48%.

**Q9:**

Melhor precisão está em precisão.

**Q10:**

Não, pois o atributo escolhido como mais relevante foi sexo, pensei que seria a classe, porém faz sentido definido ser atributo binário.

**Q11:**

Para classe no ficou precisão 76% (era 60% no OneR) , recall 98% (era 93% no OneR), já para classe yes ficou precisão 93% (era 86% no oneR) e recall 37% (no 38%OneR)

**Q12:**

O sexo, pois atributo que é binário. Classe devido a proparação de passageiros que morreram sendo da tribulação e dos que sobrevieram que são da 1° classe

**Q13:**

Dividiu em 50% dos registros sendo não sobrevivente e 50% sim.

**Q14:**

Houve pequenas mudanças, como classe 3rd 2nd e 1st aumentando e crew reduzindo.

**Q15:**

Melhorou o recall tanto para no e yes, porém pirou precisão e acurácia diminui para 72%.

**Q16:**

Nada aconteceu, ainda continua proporção 50% - 50%

**Q17:**

Os dados foram relativamente mantidos

**Q18:**

Acurácia caiu um pouco ara 71¨% (antes estava 72), precisão aumentou bem pouco (foi para 71% na média) e recall caiu para 71% na média.

**Q19:**

As técnicas de classificação estão focadas e devem ser usadas para prever a label de um atributo em questão baseado no dataset de treinamento, então ele aprende a relacionar valores de features com o valor da label da variável que será prevista. Já associação está focado em encontrar e extrair relacionamento (causa-consequência) entre os atributos/features. Basicamente dependendo do objetivo de negócio podem ser usadas juntas, pois tem objetivos diferentes, mas podem ser complementares.

DATASET IRIS - RESPOSTAS

**Q20:**

Existem 3 classes (tipos de flor) e elas estão igualmente balanceadas com 50 registros cada.

**Q21:**

Pelo gráfico aparenta que os atributos comprimento da pétala e largura da pétala conseguem separar melhor (classificar melhor) as 3 classes, pois podemos ver no gráfico cada classe mais bem dividida usando para esses 2 atributos.

**Q22:**

Os três possuem desempenho muito semelhante, o que podemos considerar como diferencial é que no SVM tem dificuldade em classificar alguns em virginica, no caso 5 eram virginica e foi classificado como versicolor. Já o Naive Bayes possui essa mesma dificuldade e também em classificar versicolor, onde 2 que eram versicolor classificou como virginica. Já árvore de decisão um registro apenas foi classificado como versicolor, porém era setosa e 2 eram virginica e foi classificado como versicolor e 3 que eram versicolor e classificou como virginica. Portanto, em números de erro de classificação entre versicolor e virginica, árvore de decisão errou levemente menos, sendo assim, considero como melhor desempenho, já que em uma base com mais registro essa diferença poderia ser maior e o J48 ter um desempenho bem mais relevante.

**Q23:**

Os três algoritmos possuem dificuldade em induzir se uma flor é versicolor ou virginica.Em ordem de compreensibilidade, ficaria:

* J48
* Naive Bayes
* SVM