Análise dos métodos supervisionados para o conjunto de dados: Titanic.

Luan C. Menezes[1], **Marcus Elias S. Freire**[1]

[1]Departamento de Ciência da Computação (DCC) – Universidade Federal da Bahia (UFBA)
Caixa Postal 40.170-110 – Salvador – BA – Brazil

{marcusfreire,luanmenezes}@dcc.ufba.br

Resumo. *O objetivo deste trabalho é analisar quais pessoas teriam maior potencial para sobreviver ao acidente do Titanic, após a utilização do conjunto de dados com alguns métodos supervisionado.*

1. Introdução

O naufrágio do RMS Titanic é um dos naufrágios mais famosos da história. A embarcação partiu em sua viagem inaugural de Southampton para Nova Iorque em 10 de abril de 1912, no caminho passando em Cherbourg-Octeville na França e por Queenstown na Irlanda. Colidiu com um *iceberg*, em 15 de abril de 1912, matando 1502 de 2224 passageiros e tripulantes.
Neste trabalho vamos analisar que tipos de pessoas provavelmente sobreviveriam a esse naufrágio.

2. Conjunto de dados

Para o conjunto de treinamento dos dados do Titanic, que foi fornecido no *site* [Kaggle 2018], com as características sobre cada passageiro, como classe social, sexo, idade, nome, quantidade de irmãos e parentes no navio, número e preço pago pelo *ticket*, assim como a cabine e o porto de embarque.

2.1. Pré-processamento

O pré-processamento foi dividido em 3 fases, na qual a primeira fase seria, conhecer a base de dados, como existência de dados faltantes, a porcentagem de cada atributo. Já na segunda fase, ocorrerá tratamento dos casos identificado na primeira fase, na terceira e última fase, os atributos serão divididos em grupos para melhorar a classificação pelos métodos.

2.1.1. Fase I: Explorando os dados

A base de dados utilizada para treinar, com 891 instâncias, está bem similar a realidade do acidente, onde o percentual de sobrevivente é de 38%. Em que, desse percentual, 81% é do sexo feminino ou pertence a primeira classe.
Em relação a falta de dados, alguns atributos têm poucas falhas, como o porto que embarcou **Embarked**, faltam em apenas 2 instâncias, mas em relação a outros atributos, como o código da Cabine **Cabin**, 77% desses dados são desconhecidos, assim como a idade **Age** que é de quase 20% de falta de dados.

Tabela 1. Instância do conjunto de dados: Titanic

Instância Número 12			
PassengerId	12	**SibSp**	0
Survived	1	**Parch**	0
Pclass	1	**Ticket**	113783
Name	Bonnell, Miss. Elizabeth	**Fare**	26.55
Sex	female	**Cabin**	C103
Age	58	**Embarked**	S

Foi verificado possíveis *outliers* em fig.1, ao gerar o gráfico *boxplot*, em relação ao preço do *Ticket*, porém, todos os preços altos que foram identificados, fora da média, eles pertenciam a primeira classe, então foi descartado a exclusão deste atributo.

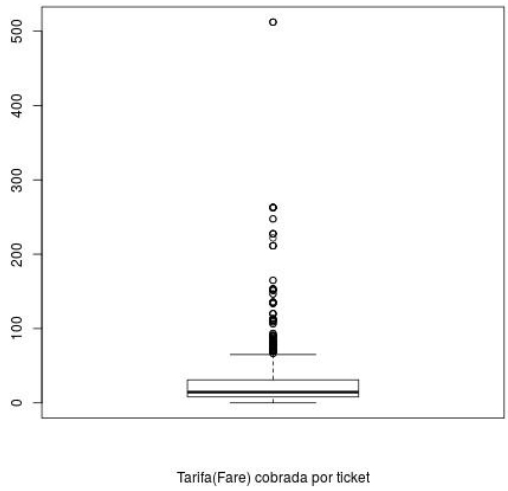

Figura 1. Possíveis outliers identificado, mas todos os pontos fora da média são da primeira classe.

2.1.2. Fase II: Tratando os dados

A eliminação manual começou com a exclusão do atributo **Name** e **PassengerId**, pois existe um nome e um id diferente para cada instância, logo, não fará diferença a exclusão desses atributos. Outro atributo que foi pensado em ser excluído, foi o código da cabine (**Cabin**), por ter tantos dados faltantes, contudo, foi visto que os que continham o código da cabine, tinham uma alta probabilidade de sobreviver, então foi útil para o método probabilístico. Assim, foi dividido em duas classes, em que continham ou não código da cabine. Já no atributo de idade, os dados faltantes foram calculados em uma interpolação entre os dados vizinhos, para que houvesse um preenchimento mais real desses dados, utilizando a biblioteca [Zeileis and Grothendieck 2005], bem como nos demais dados faltantes.

Tabela 2. Instância do conjunto de dados: Titanic, depois do tratamento

Instância Número 12								
Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Cabin	Embarked
1	1	1	1	1	1	-1	1	1

2.1.3. Fase III: Agrupando os dados

Para trabalhar com método probabilístico é necessário que os atributos, sejam categóricos, ou seja, separados em classes, então, como existiu uma classificação dos atributos, utilizamos números para representar estas classes, e assim utilizada para as redes neurais.
Como a maioria dos atributos já estavam separados em classes, como o Survived, em 0 se não sobreviveu, e 1 se sobreviveu. O Pclass, em 3 classes diferentes, 1,2 e 3, correspondente as classes sociais, foram classificados em 1,0,-1 respectivamente. O Sexo em masculino e feminino ou em 0 e 1, porto que embarcou sendo em S para Southampton, C para Cherbourg-Octeville e Q para Queenstown.
Já os atributos contínuos, foram convertidos em discretos, como foi ensinado na aula sobre árvore de decisão. Assim, o atributo Fare, ou seja, preço do Ticket, foi dividido em 3 classes, os maiores que 47 para a classe 1, o que estavam entre 10 e 47, para a classe 0 e os menores que 10 para classe -1. Mesmo modo foi feito com a idade, entretanto os limites foram menores que 20, na classe 1, entre 20 e 50 na classe 2, e os maiores que 50 na classe 3.
O objetivo era deixar todos os atributos em três classes, com os números 1,0,-1, assim poderia ser utilizado para os três métodos de classificação, todavia o **knn** mostrou ser mais eficiente, deixando os dados originais, menos a idade que foi feito o tratamento da interpolação. Pois, no cálculo da distância, para os dados contínuos, permaneceram o mesmo, mas para os dados discretos, verificava se eram diferentes, se fossem, adicionavam 10 unidades a mais na distância, para que não pudessem ser vizinhos, se fossem iguais não alteraria a distância.

3. Métodos Utilizados

Os algoritmos e técnicas utilizadas, foram retirados do livro de [Faceli et al. 2000]. Os métodos supervisionados e as técnicas de amostragem para construir a base de treinamento e teste, definindo assim um modelo para solucionar o problema de que tipo de pessoa sobreviveu ao desastre do Titanic.

3.1. Knn

Apesar de ser o método mais simples, foi o melhor desempenho, por está mais próximo do céu, de acordo com a análise ROC na figura2, além de não precisar fazer um tratamento rigoroso com a base de dados. A quantidade de rótulos acertados foi de 81,48%.

3.2. Naive Bayes

Esse método probabilístico, foi o que exigiu o tratamento mais rigoroso dos dados em que foram aproveitados para o perceptron também. Todos os atributos foram transformados em classes. Assim o melhor índice de acertos para esse método foi de 78,78%.

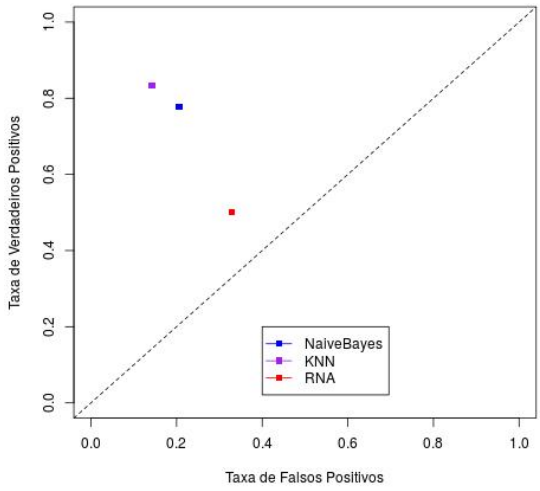

Figura 2. Análise ROC

3.3. Redes Neurais Artificiais - Perceptron

Esse método foi utilizado afim de se criar um modelo para a base de dados de treinamento, mas não foi tão preciso, pois a condição de parada seria, se não existisse mais erros ou se a quantidade de repetições fosse igual a 10.000, mas mesmo assim com o modelo que acertaria 68.35% da base de treinamento e, somente 66.32% da base de teste.

4. Conclusão

Para analisar qual tipo de pessoa sobreviveu ou não ao naufrágio, o melhor método é o modelo probabilístico, pois foi analisado os atributos que mais pesavam na escolha de um rótulo, quando indicava mais que 90% para um rótulo. Embora houvesse algum elemento de sorte envolvido na sobrevivência do naufrágio, alguns grupos de pessoas eram mais propensas a sobreviver do que outros, como mulheres, crianças e a classe alta e que possuíam o código da cabine, era de 98% para sobreviver. Já se fosse homem, da terceira classe, já o classificava com 99% a não sobreviver.

Referências

Faceli, K., Gama, J., and Lorena, A. C. (2000). *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. Grupo Gen-LTC.

Kaggle (2018). Titanic: Machine learning from disaster. `https://www.kaggle.com/c/titanic/data`. Accessed: 26-01-2018.

Zeileis, A. and Grothendieck, G. (2005). zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, 14(6):1–27.