



# Mineração de Dados

Victor H. A. Alicino

# Conteúdo

1. Base de Dados escolhida
2. Sobre o Titanic
3. Sobre a Base de Dados
4. Limpeza de Dados
5. Algoritmo Escolhido
6. Considerações Finais
7. Referências

# Base de Dados escolhida

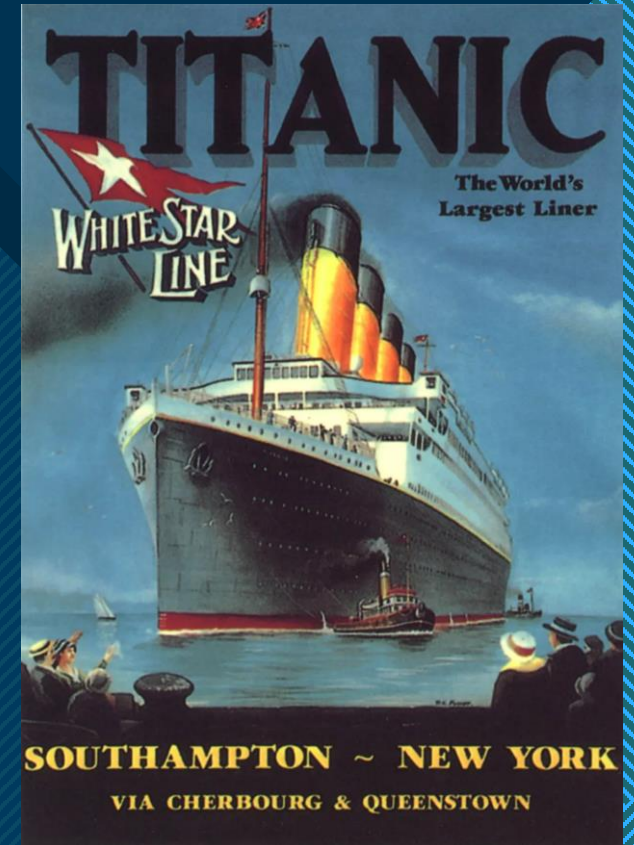
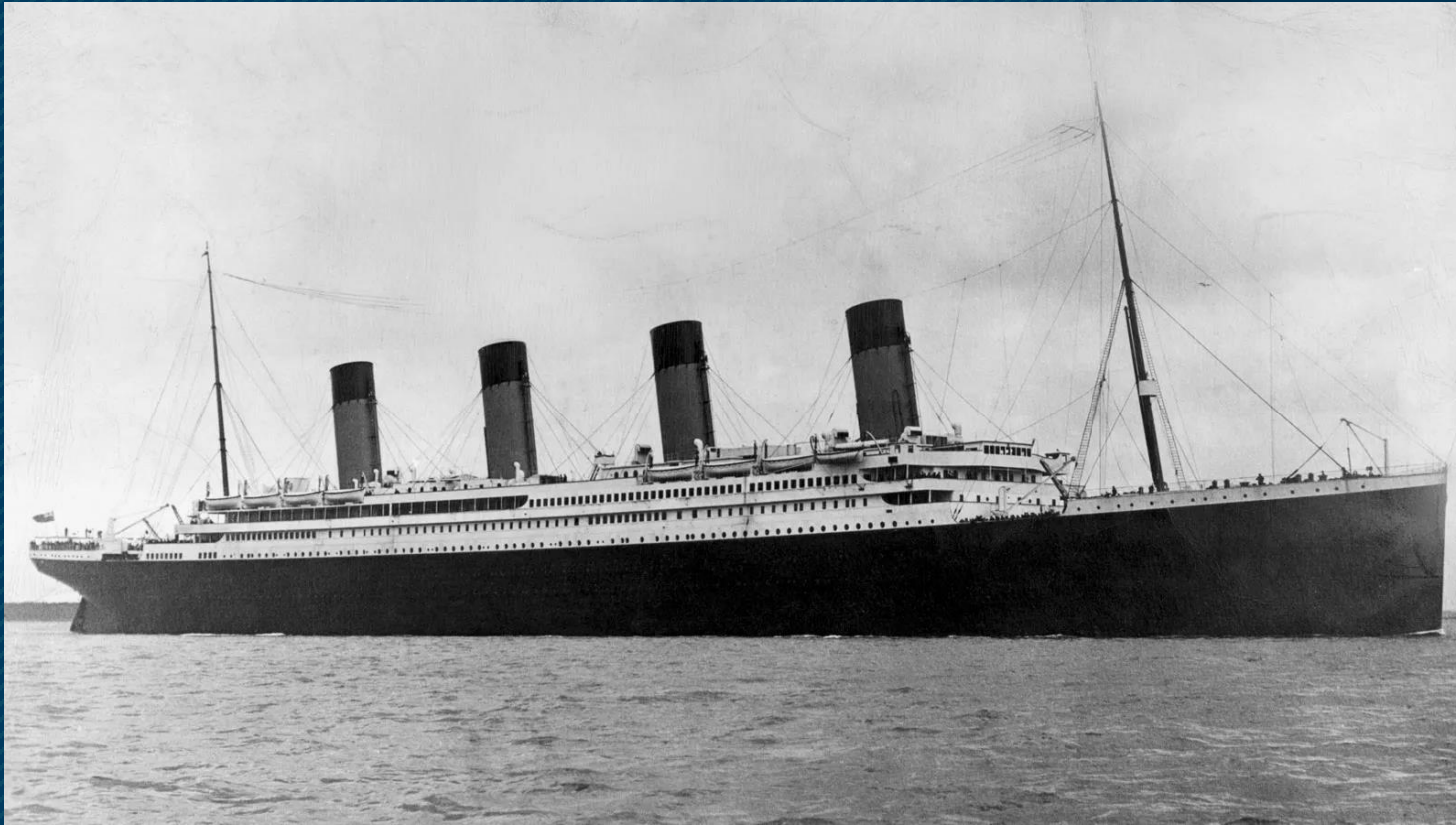
Titanic Challenge | Disponível no Kaggle

A dark, atmospheric photograph of the Titanic ship at night, with its lights glowing against the dark sea and sky. The ship is viewed from a low angle, emphasizing its massive scale.

## **Titanic - Machine Learning from Disaster**

Start here! Predict survival on the Titanic and get familiar with ML basics

# Sobre o Titanic



# Sobre a base de dados

Ela está dividida em 3 arquivos:

- train.csv
- test.csv
- gender\_submission.csv

# Sobre a base de dados

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

# TRAIN.CSV

As primeiras 9 linhas

```
PassengerId,Survived,Pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked
1,0,3,"Braund, Mr. Owen Harris",male,22,1,0,A/5 21171,7.25,S
2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs Thayer)",female,38,1,0,PC 17599,71.2833,C85,C
3,1,3,"Heikkinen, Miss. Laina",female,26,0,0,STON/O2. 3101282,7.925,S
4,1,1,"Futrelle, Mrs. Jacques Heath (Lily May Peel)",female,35,1,0,113803,53.1,C123,S
5,0,3,"Allen, Mr. William Henry",male,35,0,0,373450,8.05,S
6,0,3,"Moran, Mr. James",male,,0,0,330877,8.4583,Q
7,0,1,"McCarthy, Mr. Timothy J",male,54,0,0,17463,51.8625,E46,S
8,0,3,"Palsson, Master. Gosta Leonard",male,2,3,1,349909,21.075,S
9,1,3,"Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)",female,27,0,2,347742,11.1333,S
```

# TEST.CSV

As primeiras 9 linhas

```
PassengerId,Pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked
892,3,"Kelly, Mr. James",male,34.5,0,0,330911,7.8292,Q
893,3,"Wilkes, Mrs. James (Ellen Needs)",female,47,1,0,363272,7,S
894,2,"Myles, Mr. Thomas Francis",male,62,0,0,240276,9.6875,Q
895,3,"Wirz, Mr. Albert",male,27,0,0,315154,8.6625,S
896,3,"Hirvonen, Mrs. Alexander (Helga E Lindqvist)",female,22,1,1,3101298,12.2875,S
897,3,"Svensson, Mr. Johan Cervin",male,14,0,0,7538,9.225,S
898,3,"Connolly, Miss. Kate",female,30,0,0,330972,7.6292,Q
899,2,"Caldwell, Mr. Albert Francis",male,26,1,1,248738,29,S
900,3,"Abraham, Mrs. Joseph (Sophie Halaut Easu)",female,18,0,0,2657,7.2292,C
```



# GENDER\_SUBMISSION.CSV

As primeiras 9 linhas

PassengerId, Survived

892, 0

893, 1

894, 0

895, 0

896, 1

897, 0


898, 1

899, 0

900, 1

# Limpeza de dados

Removida as seguintes colunas



	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	725		S
3	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	712833	C85	C
4	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7925		S
5	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	531	C123	S
6	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	805		S
7	6	0	3	Moran, Mr. James	male		0	0	330877	84583		Q
8	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	518625	E46	S
9	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21075		S

# Limpeza de dados

- O arquivo “test.csv” não possuía a coluna “Survived”, esses atributos estavam presentes em “gender\_submission.csv”, porém para a realizar os testes no WEKA usando esse arquivo, foi necessário trazer essa coluna para o “test.csv”
- Também foi criado um arquivo “ext\_train.csv” que possui todas as instâncias dos arquivos “train” e “test”
- Os arquivos “train.csv”, “test.csv” e “ext\_train.csv” foram convertidos para ARFF usando o WEKA

# Limpeza de dados

```
second_step > train.arff
1  @relation train
2
3  @attribute Survived {0, 1}
4  @attribute Pclass numeric
5  @attribute Sex {male,female}
6  @attribute Age numeric
7  @attribute SibSp numeric
8  @attribute Parch numeric
9  @attribute Fare numeric
10 @attribute Embarked {S,C,Q}
```

# Tratando valores faltantes

Arquivo: train.arff

- Coluna 4 (Age) - 177 valores faltantes (20%)

Arquivo: test.arff

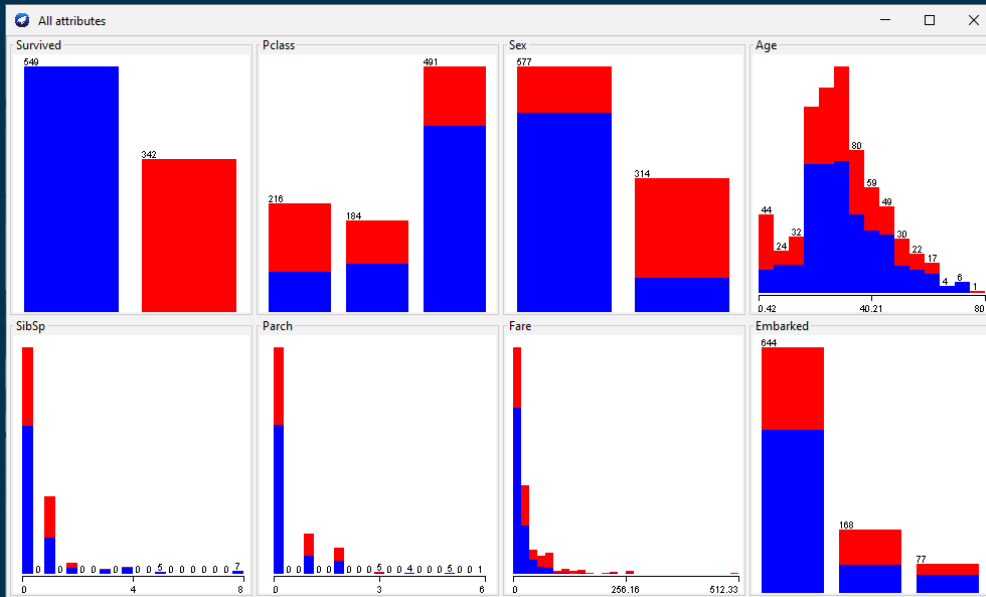
- Coluna 3 (Age) - 86 valores faltantes (20%)
- Coluna 6 (Fare) - 1 valor faltante (0%)

Arquivo: ext\_train.arff

- Coluna 4 (Age) - 263 valores faltantes (20%)
- Coluna 7 (Fare) - 1 valor faltante (0%)
- Coluna 8 (Embarked) - 1 valor faltante (0%)

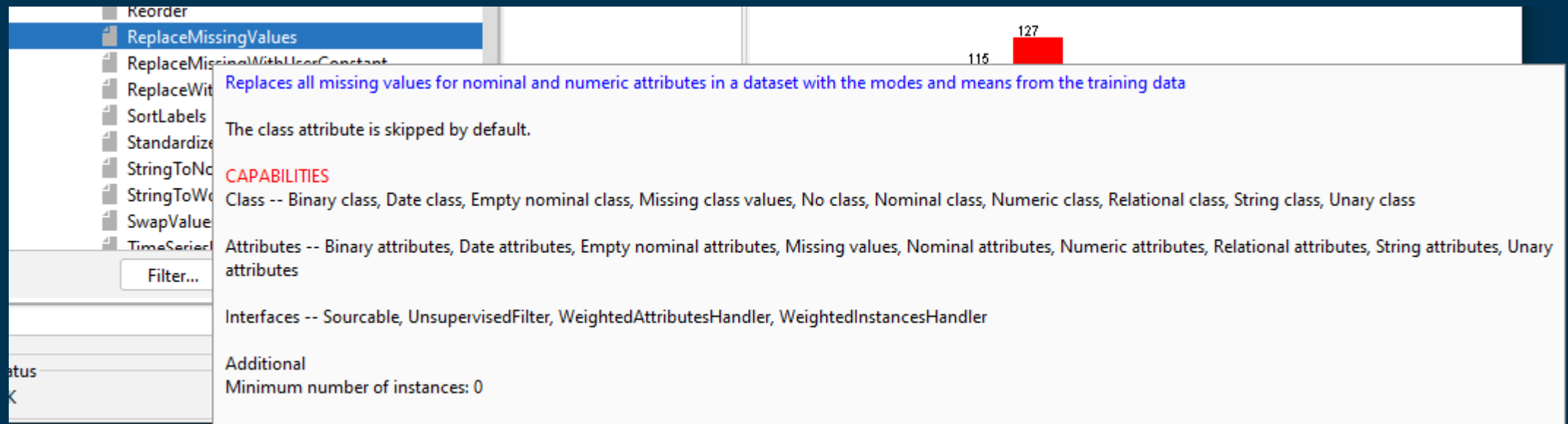
# Tratando valores faltantes

- Ocorrência dos valores faltantes não segue um padrão
- Método para tratamento escolhido: Imputação de Dados
- Método de imputação de dados escolhida: KNN



# Tratando valores faltantes

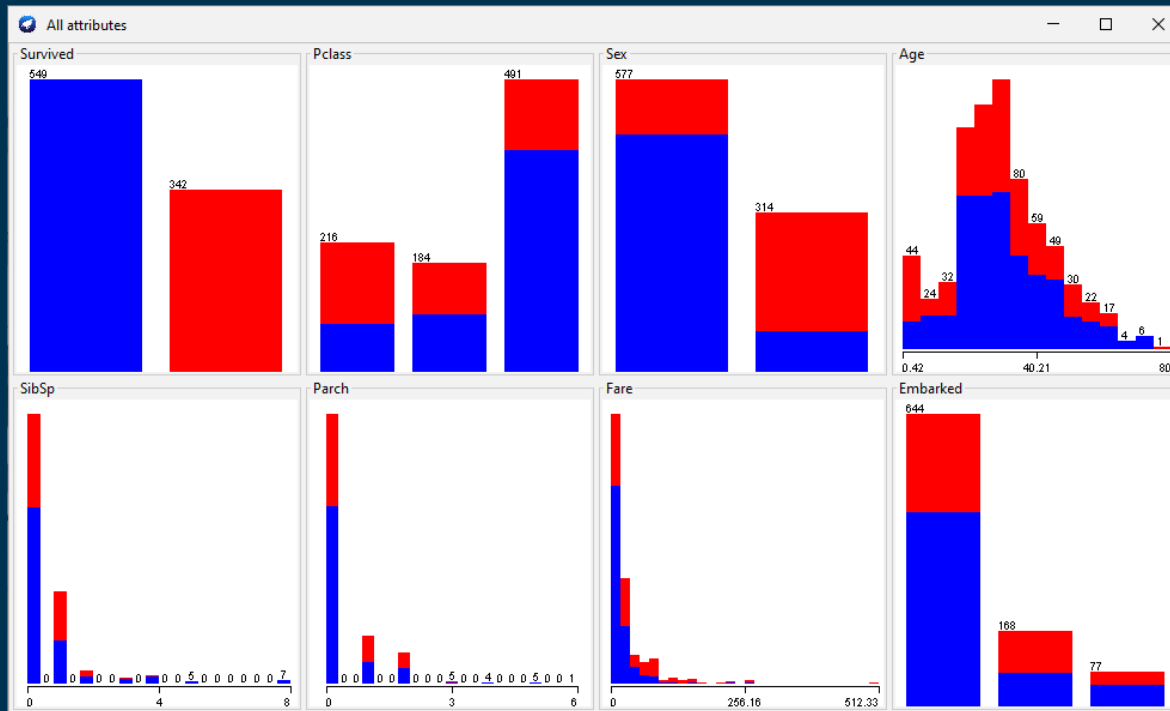
- Não encontrei imputação de dados por KNN no WEKA



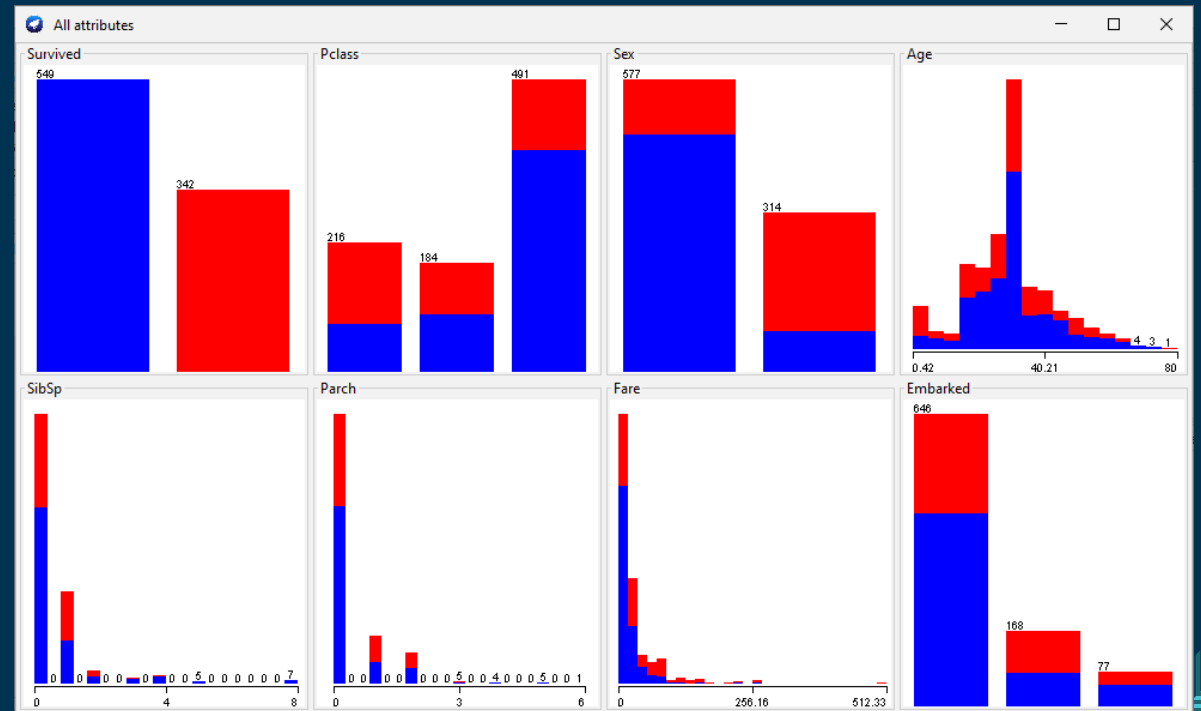
# Tratando valores faltantes

- Aplicando o ReplaceMissingValues do WEKA

Antes



Depois





# Limpeza de dados

```
third_step > train.arff
1  @relation train
2
3  @attribute Survived {0, 1}
4  @attribute Pclass numeric
5  @attribute Sex numeric
6  @attribute Age numeric
7  @attribute SibSp numeric
8  @attribute Parch numeric
9  @attribute Fare numeric
10 @attribute Embarked {S,C,Q}
```

- Convertido “Sex” para ‘numeric’

0 - Masculino

1 - Feminino

# Algoritmo escolhido

- J48
- Árvore de Decisão
  - Fácil interpretação
  - A base contém dados categóricos

# Caso base

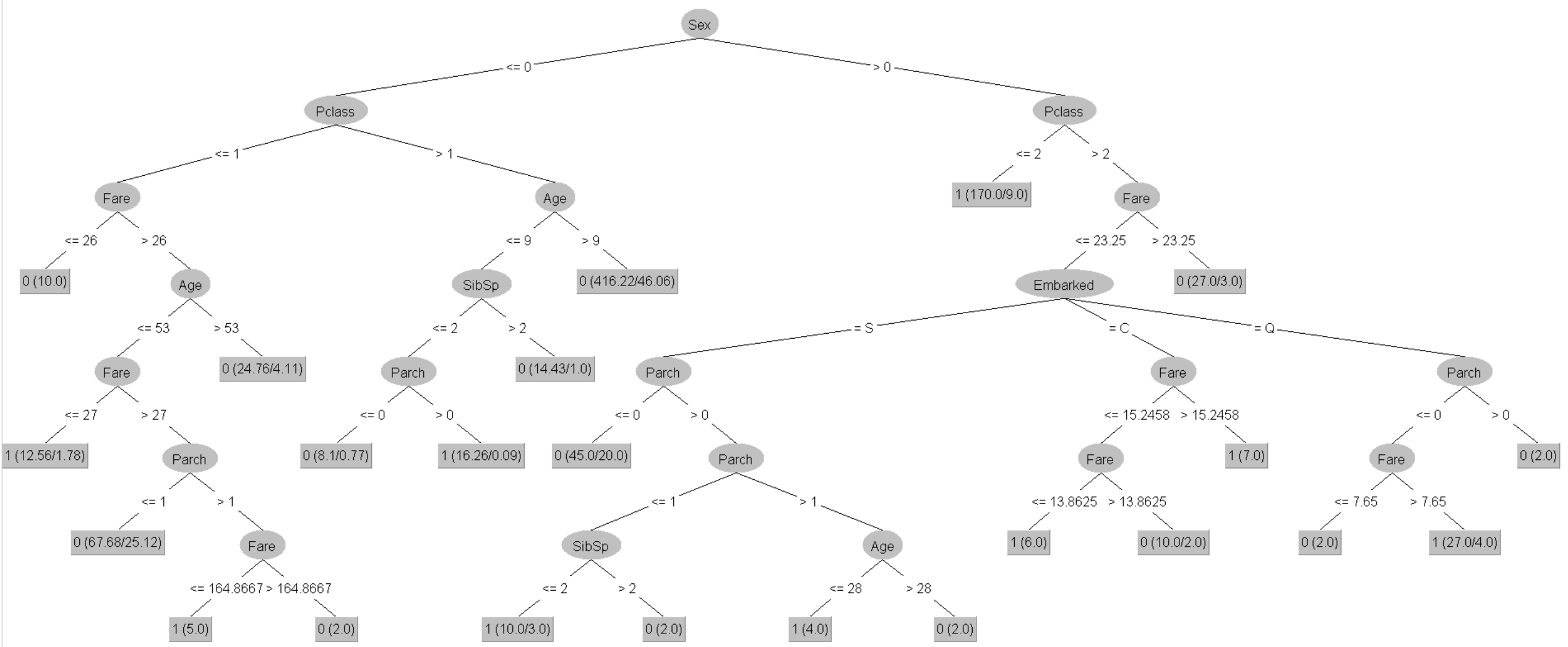
train.arff | J48 | Testado com test.arff

Precisão: 87,4%

Falsos Positivos: 18%

Matriz de Confusão

Pereceu	Sobreviveu
251	15
38	114



# Removendo o atributo 'Embarked'

train\_minus\_embarked.arff | J48 | Testado com test.arff

**Precisão: 95,1%**

**Falsos Positivos: 4,6%**

Matriz de Confusão

Pereceu	Sobreviveu
251	15
6	146

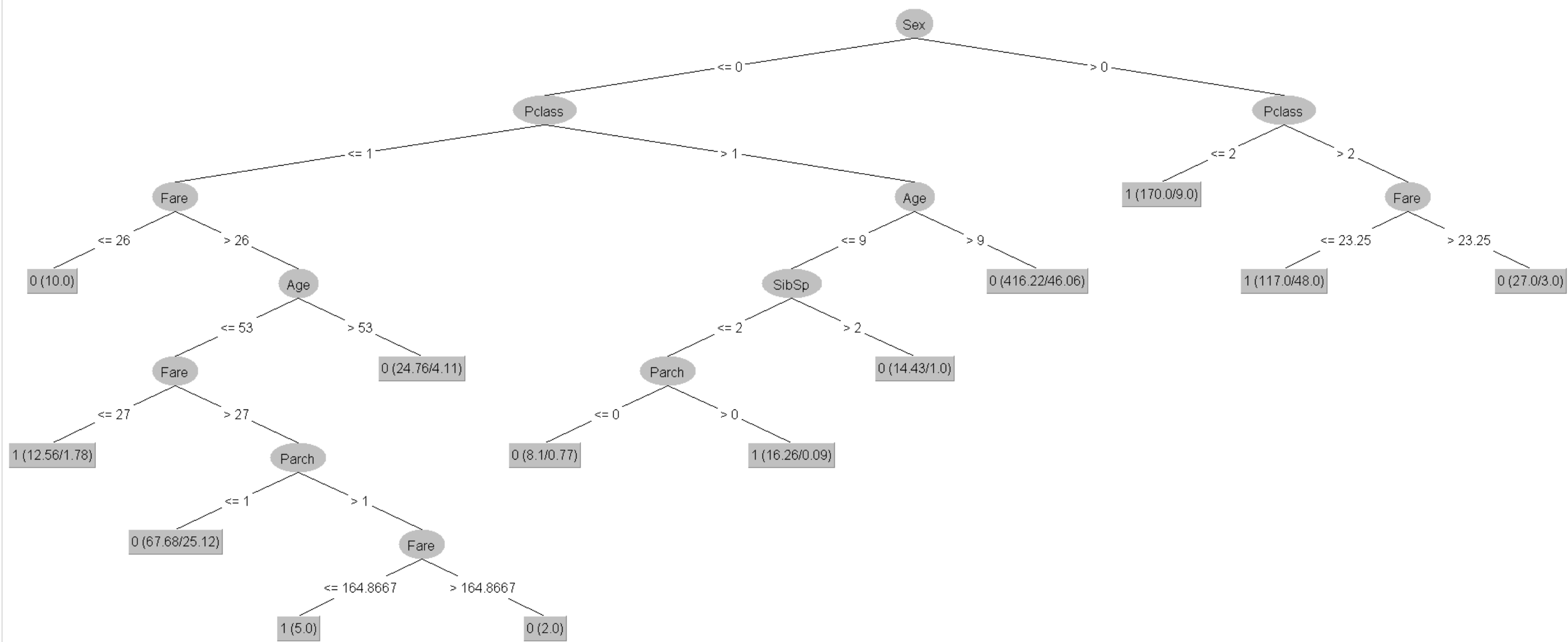
train\_minus\_embarked.arff | J48 | Testado com cross-validation

**Precisão: 81,93%**

**Falsos Positivos: 22,3%**

Matriz de Confusão

Pereceu	Sobreviveu
488	61
100	242



# Testando com apenas 4 atributos

train\_minus\_embarked.arff | J48 | Testado com test.arff

**Precisão: 93,3%**

**Falsos Positivos: 10,9%**

Matriz de Confusão

Pereceu	Sobreviveu
263	15
25	127

train\_minus\_embarked.arff | J48 | Testado com cross-validation

**Precisão: 79,1%**

**Falsos Positivos: 25%**

Matriz de Confusão

Pereceu	Sobreviveu
475	74
110	232

# Testando com apenas 4 atributos

Possível overfitting

train\_minus\_embarked.arff | J48 | Testado com test.arff

**Precisão: 93,3%**

**Falsos Positivos: 10,9%**

Matriz de Confusão

Pereceu	Sobreviveu
263	15
25	127

train\_minus\_embarked.arff | J48 | Testado com cross-validation

**Precisão: 79,1%**

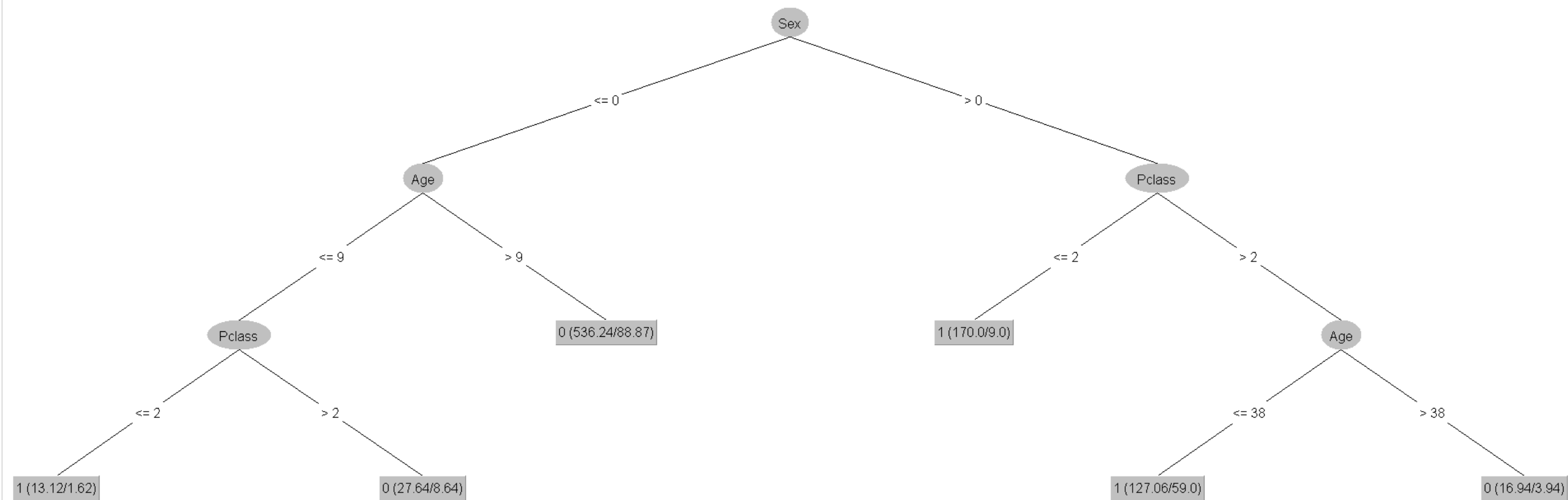
**Falsos Positivos: 25%**

Matriz de Confusão

Pereceu	Sobreviveu
475	74
110	232



Captura Retangular



# Comparando com a realidade

- Birkenhead Drill
  - “Mulheres e crianças primeiro”
- 72% das mulheres sobreviveram
- 50% das crianças sobreviveram
- Apenas 16% dos homens sobreviveram

# Dificuldades encontradas

- Não conhecer as implementações dos algoritmos no WEKA
- Não conseguir realizar a imputação de dados da forma que gostaria no WEKA

# Considerações finais

- Apesar da alta possibilidade do modelo com a melhor precisão aqui mostrado (95%) estar apresentando overfit, o mesmo apresenta 80% de precisão nos dados reais do acidente

Weka Explorer

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

ChooseJ48 -C 0.25 -M 2

Test options

Use training set

Supplied test set

Cross-validation

Percentage split

Set...

Folds10

%66

More options...

(Nom) Survived

StartStop

Result list (right-click for options)

20:52:07 - trees.J48

Classifier output

Kappa statistic0.8928

Mean absolute error0.2055

Root mean squared error0.2942

Relative absolute error43.8686 %

Root relative squared error61.0943 %

Total Number of Instances418

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,944	0,039	0,977	0,944	0,960	0,894	0,947	0,978	0
	0,961	0,056	0,907	0,961	0,933	0,894	0,947	0,890	1
Weighted Avg.	0,950	0,046	0,951	0,950	0,950	0,894	0,947	0,946	

=== Confusion Matrix ===

a   b   <-- classified as

251 15 |   a = 0

6 146 |   b = 1

=== Re-evaluation on test set ===

User supplied test set

Relation:    titanic\_real

Instances:    unknown (yet). Reading incrementally

Attributes:   8

=== Summary ===

Correctly Classified Instances178080.6525 %

Incorrectly Classified Instances42719.3475 %

Kappa statistic0.5215

Mean absolute error0.2771

Root mean squared error0.3842

Total Number of Instances2207

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,924	0,442	0,815	0,924	0,866	0,535	0,795	0,842	0
	0,558	0,076	0,778	0,558	0,650	0,535	0,795	0,680	1
Weighted Avg.	0,807	0,324	0,803	0,807	0,797	0,535	0,795	0,790	


=== Confusion Matrix ===

a   b   <-- classified as

1383 113 |   a = 0

314 397 |   b = 1

StatusOK

Log x 0

# Referências

**Titanic: Demographics of the Passengers.** Disponível em: <<http://www.icyousee.org/titanic.html>>.

**HMS Birkenhead and the Birkenhead Drill - Women and Children First.** Disponível em: <<https://www.historic-uk.com/CultureUK/Women-Children-First/>>.

**Titanic Survivors.** Disponível em: <<https://titanicfacts.net/titanic-survivors/>>.

**Titanic Passengers and Crew Complete List.** Disponível em: <<https://www.kaggle.com/datasets/aliaamiri/titanic-passengers-and-crew-complete-list>>.

**Titanic - Machine Learning from Disaster.** Disponível em: <<https://www.kaggle.com/competitions/titanic>>.

The background features a dark blue field on the right and a light blue field on the left, separated by a diagonal line. A thin, dark blue line runs parallel to the diagonal, and a thin, light blue line with horizontal stripes runs parallel to the dark blue line.

Obrigado