U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Inteligência Artificial(CC2006)

Relatório do Projeto Prático 2

Manuel Sá - up201805273
Miguel Ramos - up201805242

Introdução

As Árvores de Decisão são um dos modelos mais práticos e mais usados em inferência indutiva. Este método representa funções como árvores de decisão. Estas árvores são treinadas de acordo com um conjunto de treino (exemplos previamente classificados) e posteriormente, outros exemplos são classificados de acordo com essa mesma árvore. Para a construção destas árvores são usados algoritmos como o ID3, ASSISTANT e C4.5.

Dados iniciais

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
845	0	3	Culumovic, Mr. J	male	17	0	0	315090	8.6625		S
569	0	3	Doharr, Mr. Tann	male	28(média)	0	0	2686	7.2292		C
792	0	2	Gaskell, Mr. Alfre	male	16	0	0	239865	26		S
752	1	3	Moor, Master. M	male	6	0	1	392096	12.475	E121	S
40	1	3	Nicola-Yarred, M	female	14	1	0	2651	11.2417		C
487	1	1	Hoyt, Mrs. Frede	female	35	1	0	19943	90	C93	S
852	0	3	Svensson, Mr. Jo	male	74	0	0	347060	7.775		S
814	0	3	Andersson, Miss.	female	6	4	2	347082	31.275		S
164	0	3	Calic, Mr. Jovo	male	17	0	0	315093	8.6625		S
720	0	3	Johnson, Mr. Ma	male	33	0	0	347062	7.775		S
170	0	3	Ling, Mr. Lee	male	28	0	0	1601	56.4958		S
858	1	1	Daly, Mr. Peter D	male	51	0	0	113055	26.55	E17	S
390	1	2	Lehmann, Miss.	female	17	0	0	SC 1748	12		C
697	0	3	Kelly, Mr. James	male	44	0	0	363592	08.05		S
360	1	3	Mockler, Miss. H	female	27(média)	0	0	330980	7.8792		Q
608	1	1	Daniel, Mr. Robe	male	27	0	0	113804	30.5		S
671	1	2	Brown, Mrs. Tho	female	40	1	1	29750	39		S
641	0	3	Jensen, Mr. Hans	male	20	0	0	350050	7.8542		S
527	1	2	Ridsdale, Miss. L	female	50	0	0	W./C. 14258	10.5		S
349	1	3	Coutts, Master. V	male	3	1	1	C.A. 37671	15.9		S
					(74-3)/4 = 17.75						
					Média M: 28						
					Média F: 27						

Como pedido no enunciado, as colunas Ticket e Cabin não foram consideradas para a criação das árvores de decisão. Da mesma forma, também não foram consideradas as tabelas PassengerId e Name, porque são atributos únicos que não importam ao problema.

Impureza de Gini

Para escolhermos a raíz da nossa árvore, calculamos a Impureza de Gini para cada uma das colunas, identificando-as como classes. Essas classes serão depois inseridas na árvore, de acordo com qual tem menor impureza. Para calcular o valor da impureza a cada nível, são utilizadas duas fórmulas:

A primeira fórmula é a seguinte:

1-(ns/nt)^2-(nm/nt)^2

ns - número de pessoas sobreviventes de nt;
nm - número de pessoas que morreram de nt;
nt - número total de pessoas na subclasse.

Este cálculo permite-nos obter a impureza dos dados para cada subclasse, por exemplo, se escolhermos a classe Sexo(Homem) terá duas subclasses, dependendo se a resposta a se uma pessoa pertence à classe for Sim ou Não. O número de subclasses é sempre 2, devido ao resultado binário sim/não.

Quando temos ambas as impurezas, podemos então calcular a Impureza de Gini da classe. A fórmula é a seguinte:

(p/t)*igp + (np/t)*ignp

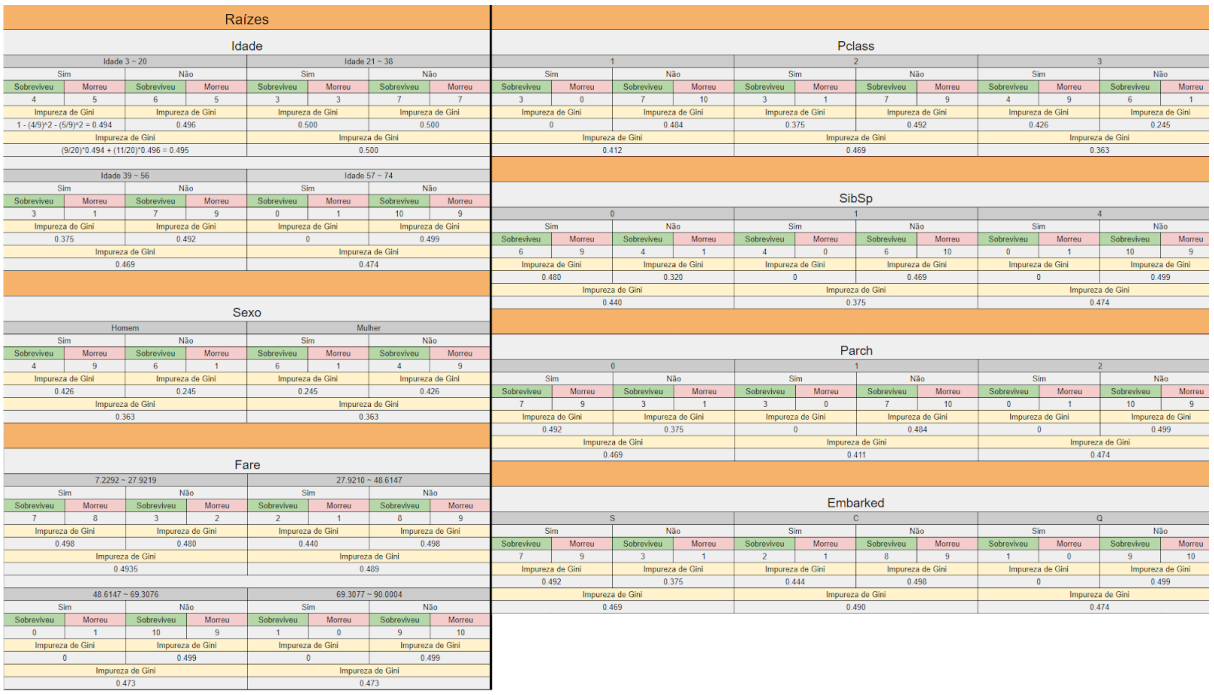
t - número total de pessoas pertencentes à classe;
p - pessoas de t pertencentes à subclasse Sim;
np - pessoas de t não pertencentes à subclasse Não;
igp - Impureza de Gini de p;
ignp - Impureza de Gini de np.

Para ajuda à construção das tabelas onde guardamos os valores, escrevemos o seguinte código em javascript, onde GINISOLO calcula a impureza da subclasse, e GINIDOUBLE para a classe toda.

```
function GINISOLO(input1,input2){
  var firstPart = (input1/(input1+input2))**2;
  var secondPart = (input2/(input1+input2))**2;
  return 1 - firstPart - secondPart;
}

function GINIDOUBLE(input1,input2,resultado1e2,input3,input4,resultado3e4) {
  return ((input1 + input2)/(input1+input2+input3+input4))*resultado1e2 +
((input3 + input4)/(input1+input2+input3+input4))*resultado3e4;
}
```

Depois de calculadas as impurezas de cada classe, resta apenas escolher a que tem menor valor e defini-la como raíz. Neste caso, as classes com menor impureza foram a Sexo(Homem) e Sexo(Mulher) e são iguais, visto os dados serem inversos entre elas. Decidimos escolher Sexo(Homem como a nossa raíz, que irá dividir a população entre 13 (Homens) e 7 (Mulheres) pessoas, para as respostas Sim e Não respetivamente.



Raízes

Idade

Idade 3 ~ 20				Idade 21 ~ 38			
Sim		Não		Sim		Não	
Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu
4	5	6	5	3	3	7	7
Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini	
1 - (4/9)^2 - (5/9)^2 = 0.494		0.496		0.500		0.500	
Impureza de Gini				Impureza de Gini			
(9/20)*0.494 + (11/20)*0.496 = 0.495				0.500			

Idade 39 ~ 56				Idade 57 ~ 74			
Sim		Não		Sim		Não	
Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu
3	1	7	9	0	1	10	9
Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini	
0.375		0.492		0		0.499	
Impureza de Gini				Impureza de Gini			
0.469				0.474			

Sexo

Homem				Mulher			
Sim		Não		Sim		Não	
Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu
4	9	6	1	6	1	4	9
Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini	
0.426		0.245		0.245		0.426	
Impureza de Gini				Impureza de Gini			
0.363				0.363			

Fare

7.2292 ~ 27.9219				27.9210 ~ 48.6147			
Sim		Não		Sim		Não	
Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu
7	8	3	2	2	1	8	9
Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini	
0.498		0.480		0.440		0.498	
Impureza de Gini				Impureza de Gini			
0.4935				0.489			

48.6147 ~ 69.3076				69.3077 ~ 90.0004			
Sim		Não		Sim		Não	
Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu
0	1	10	9	1	0	9	10
Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini	
0		0.499		0		0.499	
Impureza de Gini				Impureza de Gini			
0.473				0.473			

Pclass

1				2				3			
Sim		Não		Sim		Não		Sim		Não	
Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu
3	0	7	10	3	1	7	9	4	9	6	1
Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini	
0		0.484		0.375		0.492		0.426		0.245	
Impureza de Gini				Impureza de Gini				Impureza de Gini			
0.412				0.469				0.363			

SibSp

0				1				4			
Sim		Não		Sim		Não		Sim		Não	
Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu
6	9	4	1	4	0	6	10	0	1	10	9
Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini	
0.480		0.320		0		0.469		0		0.499	
Impureza de Gini				Impureza de Gini				Impureza de Gini			
0.440				0.375				0.474			

Parch

0				1				2			
Sim		Não		Sim		Não		Sim		Não	
Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu
7	9	3	1	3	0	7	10	0	1	10	9
Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini	
0.492		0.375		0		0.484		0		0.499	
Impureza de Gini				Impureza de Gini				Impureza de Gini			
0.469				0.411				0.474			

Embarked

S				C				Q			
Sim		Não		Sim		Não		Sim		Não	
Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu
7	9	3	1	2	1	8	9	1	0	9	10
Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini	
0.492		0.375		0.444		0.498		0		0.499	
Impureza de Gini				Impureza de Gini				Impureza de Gini			
0.469				0.490				0.474			

Tendo agora uma raiz, será necessário traçar os dois caminhos (caso pertença à semi tabela da raiz ou não). Para cada lado o método é o mesmo: ver o número de pessoas que pertencem e não pertencem e considerar esse valor como o novo total (ex: se total = 20 e apenas 13 pessoas pertencem à subclasse da raiz, 13 será o novo total de pessoas no caminho de ‘sim’). O processo para decidir o novo nó é o mesmo da raíz, excluindo sempre as classes já consideradas como nós internos. Para este caso, temos que a classe com menor impureza é a PClass(1), e passará então a ser o próximo nó da árvore. Como o caminho ‘Sim’ leva a todos os membros da subclasse sobreviverem, essa parte da árvore termina nesse nó e a população do caminho ‘Não’ continua a árvore. Repetimos este processo enquanto existirem classes por inserir na árvore ou até todos os caminhos da árvore levares a um “Sobrevive” ou “Morre”.

Raíz: Homem,Sim

Idade

Idade 3 ~ 20				Idade 21 ~ 38			
Sim		Não		Sim		Não	
Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu
2	4	2	5	1	3	3	6
Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini	
0,4444444444		0,4081632653		0,375		0,4444444444	
Impureza de Gini				Impureza de Gini			
0,4249084249				0,4230769231			

Idade 39 ~ 56				Idade 57 ~ 74			
Sim		Não		Sim		Não	
Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu
1	1	3	8	0	1	4	8
Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini	
0,5		0,3966942149		0		0,4444444444	
Impureza de Gini				Impureza de Gini			
0,4125874126				0,4102564103			

Fare

7.2292 ~ 27.9219				27.9210 ~ 48.6147			
Sim		Não		Sim		Não	
Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu
3	9	1	0	1	0	3	9
Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini	
0,375		0		0		0,375	
Impureza de Gini				Impureza de Gini			
0,3461538462				0,3461538462			

48.6147 ~ 69.3076				69.3077 ~ 90.0004			
Sim		Não		Sim		Não	
Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu
0	1	4	8	0	0	4	9
Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini	
0		0,4444444444		#NUM!		0,426035503	
Impureza de Gini				Impureza de Gini			
0,4102564103				#NUM!			

Parch

0				1				2			
Sim		Não		Sim		Não		Sim		Não	
Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu
2	9	2	0	2	0	2	9	0	0	4	9
Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini	
0,2975206612		0		0		0,2975206612		#NUM!		0,426035503	
Impureza de Gini				Impureza de Gini				Impureza de Gini			
0,2517482517				0,2517482517				#NUM!			

Embarked

S				C				Q			
Sim		Não		Sim		Não		Sim		Não	
Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu
4	8	0	1	0	1	4	8	0	0	0	0
Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini	
0,4444444444		0		0		0,4444444444		#NUM!		#NUM!	
Impureza de Gini				Impureza de Gini				Impureza de Gini			
0,4102564103				0,4102564103				#NUM!			

Pclass

1				2				3			
Sim		Não		Sim		Não		Sim		Não	
Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu
2	0	2	9	0	1	4	8	2	8	2	1
Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini	
0		0,2975206612		0		0,4444444444		0,32		0,4444444444	
Impureza de Gini				Impureza de Gini				Impureza de Gini			
0,2517482517				0,4102564103				0,3487179487			

SibSp

0				1				4			
Sim		Não		Sim		Não		Sim		Não	
Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu
3	9	1	0	1	0	3	9	0	0	4	9
Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini	
0,375		0		0		0,375		#NUM!		0,426035503	
Impureza de Gini				Impureza de Gini				Impureza de Gini			
0,3461538462				0,3461538462				#NUM!			

Raíz: Homem,Sim -> Pclass(1),Não

Idade

Idade 3 ~ 20				Idade 21 ~ 38			
Sim		Não		Sim		Não	
Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu
2	3	0	5	0	3	2	6
Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini	
0,48		0		0		0,375	
Impureza de Gini				Impureza de Gini			
0,24				0,2727272727			

Idade 39 ~ 56				Idade 57 ~ 74			
Sim		Não		Sim		Não	
Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu
0	1	2	8	0	1	2	8
Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini	
0		0,32		0		0,32	
Impureza de Gini				Impureza de Gini			
0,2909090909				0,2909090909			

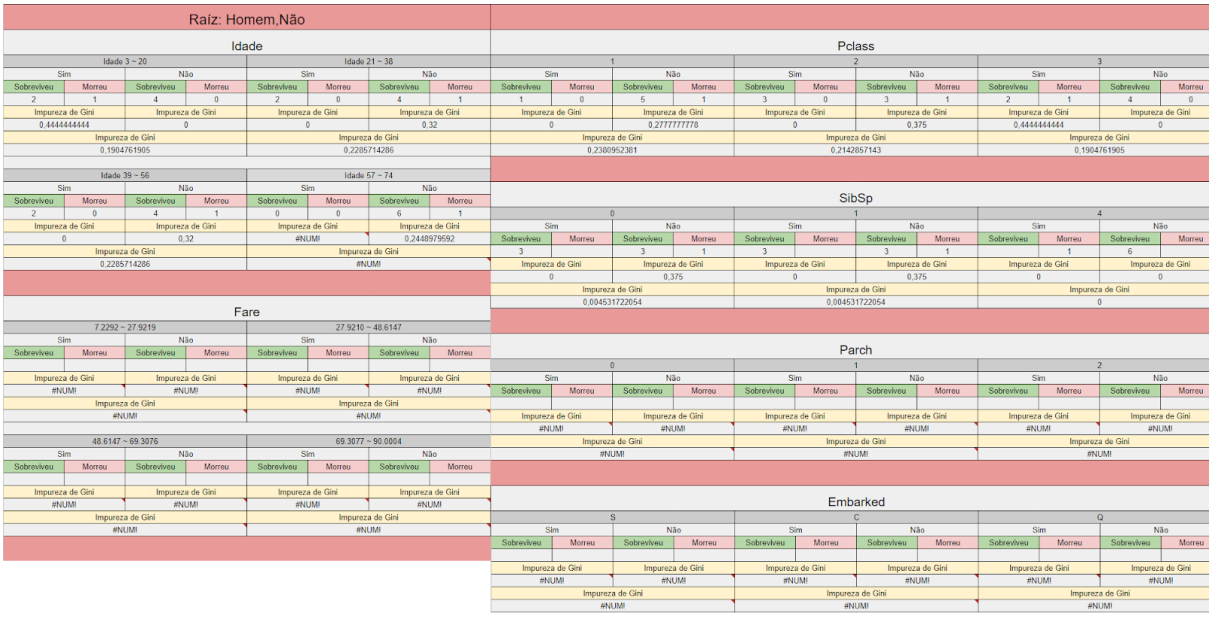
SibSp

0				1				4			
Sim		Não		Sim		Não		Sim		Não	
Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu
1	9	1	0	1	0	1	9	0	0	2	9
Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini	
0,18		0		0		0,18		#NUM!		0,2975206612	
Impureza de Gini				Impureza de Gini				Impureza de Gini			
0,1636363636				0,1636363636				#NUM!			

Parch

0				1				2			
Sim		Não		Sim		Não		Sim		Não	
Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu	Sobreviveu	Morreu
0	9	2	0	2	0	0	9	0	0	2	9
Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini		Impureza de Gini	
0		0		0		0		#NUM!		0,2975206612	
Impureza de Gini				Impureza de Gini				Impureza de Gini			
0				0				#NUM!			

Neste caso, mal encontramos uma classe com Impureza de Gini igual a 0 em SibSp(4), significa que divide perfeitamente a população (sendo que só restava 1 morto), mal descoberta não foi necessário calcular o resto das impurezas e inserimos este último nó na tabela.



Dadas todas as informações necessárias, obtemos a seguinte árvore de decisão:

Ganho de Informação

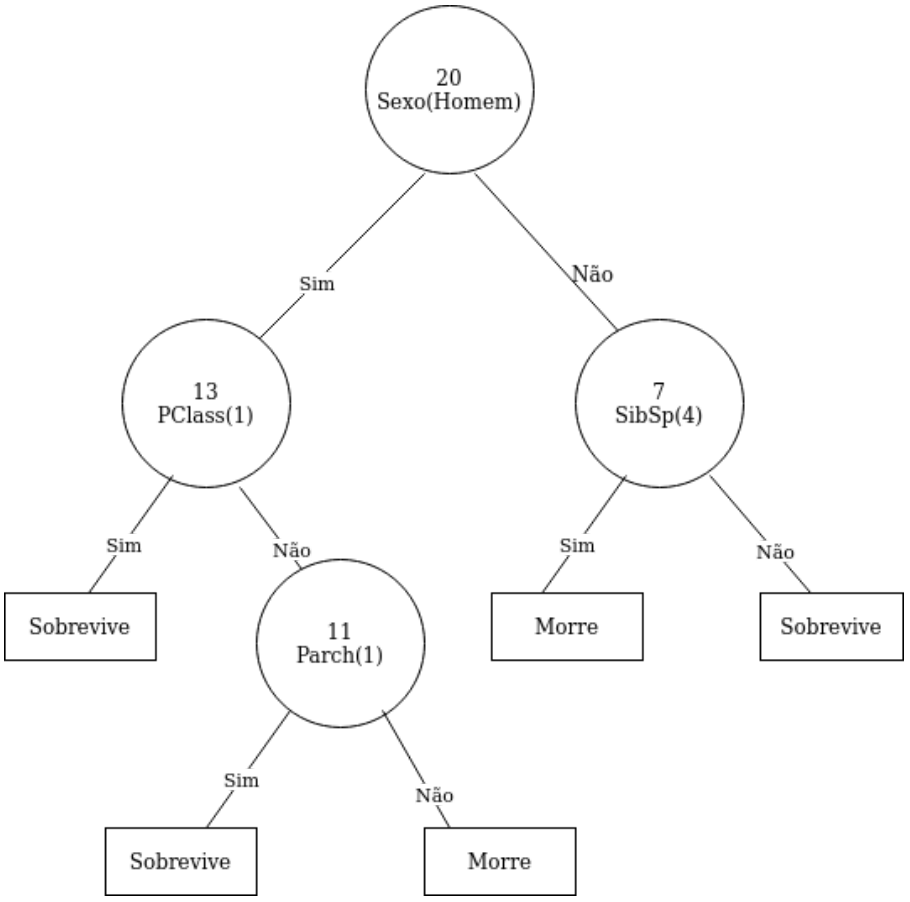
Repetindo o processo mas agora para o Ganho de Informação, voltamos a definir classes e subclasses de forma a calcular os ganhos e entropias das mesmas. Os valores são calculados das seguintes formas:

t total de pessoas;
fa frequência absoluta- número de pessoas que pertencem à subclasse;
s - número de pessoas da frequência absoluta que sobrevivem;
m - número de pessoas da frequência absoluta que morreram;
i - numerador da subclasse
k - indicador da última subclasse

$$E(conjunto) = -(s/t) * log2(s/t) - (m/t) * log2(m/t)$$

$$GI = E(conjunto) - \sum_{i=1}^{k} (fa(i)/t) * E(i)$$

Para descobrir a raiz da árvore, criámos tabelas para todas as classes. A raiz da nossa árvore será a subclasse com menor Entropia da classe com maior Ganho de informação. O ganho de



informação de cada é calculado com classe é calculado com base na entropia do conjunto a que pertence, por exemplo, numa população de 13 pessoas, t = 13.
Caso existam duas subclasses de entropia igual, selecionamos a que tem maior frequência absoluta. Se a entropia da subclasse for 0, quer dizer que o nó irá gerar uma folha do tipo "sobrevive" ou "morre", dividindo bem a população, e o processo é repetido até não existirem mais classes para explorar ou toda a população acabar inserida em ramos "sobrevive" ou "morre".

Para o cálculo das entropias, voltamos a escrever um pequeno código em javascript para agilizar o processo:

```
function ENTROPY(freq, yes, no) {
  var entropy;
  if(yes == 0 || no == 0) return 0;
  entropy = -(yes/freq)*(Math.log(yes/freq)/Math.log(2)) -
(no/freq)*(Math.log(no/freq)/Math.log(2));
  return entropy;
}
```

Entropia do conjunto:	-(10/20)*log2(10/20) - (10/20)*log2(10/20) = 1

Raízes

Valores (Idade)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Informação
3~20	9	4	5	0,9910760598	1 - ((9/20)*0.991 + (6/20)*1 + (4/20)*0.811 + (1/20)*0) = 0.09185
21~38	6	3	3	1	
39~56	4	3	1	0,8112781245	
57~74	1	0	1	0	

Valores (Sexo)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
Homem	13	4	9	0,8904916402	0.21465
Mulher	7	6	1	0,5916727786	

Valores (Pclass)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
1	3	3	0	0	0.2593
2	4	3	1	0,8112781245	
3	13	4	9	0,8904916402	

Valores (SibSp)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
0	15	6	9	0,9709505945	0.2725
1	4	4	0	0	
4	1	0	1	0	

Valores (Parch)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
0	16	7	9	0,9886994083	0.2096
1	3	3	0	0	
2	1	0	1	0	

Valores (Fare)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
7.2292 ~ 27.9219	15	7	8	0,996791632	0.1153
27.9210 ~ 48.6147	3	2	1	0,9182958341	
48.6147 ~ 69.3076	1	0	1	0	
69.3077 ~ 90.0004	1	1	0	0	

Valores (Embarked)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
S	16	7	9	0,9886994083	0.0711
C	3	2	1	0,9182958341	
Q	1	1	0	0	

Entropia do conjunto:	-(6/16)*log2(6/16) - (10/16)*log2(10/16) = 0.954434

Raíz: SibSp(1) Não

Valores (Idade)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
3~20	7	2	5	0,8631205686	0.101622
21~38	5	2	3	0,9709505945	
39~56	3	2	1	0,9182958341	
57~74	1	0	1	0	

Valores (Sexo)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
Homem	12	3	9	0,8112781245	0.143164
Mulher	4	3	1	0,8112781245	

Valores (Pclass)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
1	2	2	0	0	0.312034
2	3	2	1	0,9182958341	
3	11	2	9	0,6840384356	

Valores (Parch)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
0	14	5	9	0,9402859587	0.130184
1	1	1	0	0	
2	1	0	1	0	

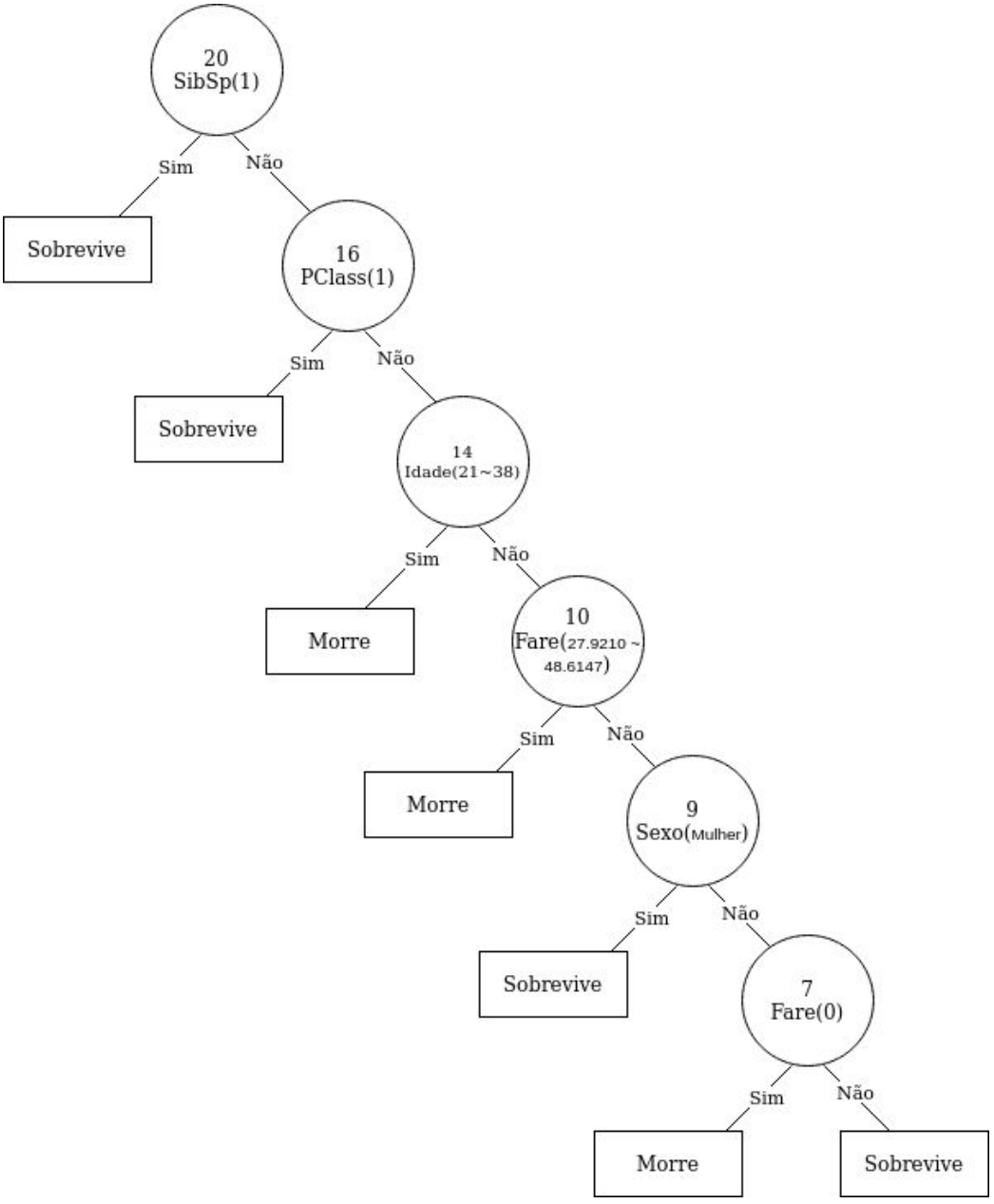
Entropia do conjunto:	-(4/14)*log2(4/14) - (10/14)*log2(10/14) = 0.863120				
Raíz: SibSp(1) Não -> PClass(1) Não					
Valores (Idade)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
3~20	7	2	5	0,8631205686	0.43156
21~38	4	0	4	0	
39~56	2	1	1	1	
57~74	1	0	1	0	
Valores (Sexo)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
Homem	10	1	9	0,4689955936	0.296332
Mulher	4	3	1	0,8112781245	
Valores (Parch)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
0	12	2	10	0,6500224216	0.305958
1	1	1	0	0	
2	1	0	1	0	
Valores (Fare)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
7.2292 ~ 27.9219	12	4	8	0,9182958341	0.07601
27.9210 ~ 48.6147	1	0	1	0	
48.6147 ~ 69.3076	1	0	1	0	
Valores (Embarked)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
S	11	2	9	0,6840384356	0.276809
C	2	1	1	1	
Q	1	1	0	0	

Entropia do conjunto:	-(4/10)*log2(4/10) - (6/10)*log2(6/10) = 0.970951				
Raíz: SibSp(1) Não -> PClass(1) Não -> Idade(21~38) Não					
Valores (Sexo)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
Homem	7	1	6	0,5916727786	0.281292
Mulher	3	2	1	0,9182958341	
Valores (Parch)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
0	8	2	6	0,8112781245	0.321935
1	1	1	0	0	
2	1	0	1	0	
Valores (Fare)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
7.2292 ~ 27.9219	9	1	8	0,5032583348	0.518019
27.9210 ~ 48.6147	1	0	1	0	
Valores (Embarked)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
S	9	2	7	0,7642045065	0.283167
C	1	1	0	0	

Entropia do conjunto:	-(4/9)*log2(4/9) - (5/9)*log2(5/9) = 0.991076				
Raíz: SibSp(1) Não -> PClass(1) Não -> Idade(21~38) Não -> Fare(27.9210 ~ 48.6147) Não					
Valores (Sexo)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
Homem	7	1	6	0,5916727786	0.530888
Mulher	2	2	0	0	
Valores (Parch)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
0	8	2	6	0,8112781245	0.26994
1	1	1	0	0	
Valores (Embarked)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
S	8	2	6	0,8112781245	0.26994
C	1	1	0	0	

Entropia do conjunto:	-(2/7)*log2(2/7) - (5/7)*log2(5/7) = 0.863121				
Raíz: SibSp(1) Não -> PClass(1) Não -> Idade(21~38) Não -> Fare(27.9210 ~ 48.6147) Não -> Sexo(Mulher) Não					
Valores (Parch)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
0	6	0	6	0	
1	1	1	0	0	
Valores (Embarked)	Frequência relati	sobreviveram	morreram	entropia	Ganho de Inf.
S	7	1	6	0,5916727786	0.271448

Dadas todas as informações necessárias, obtemos a seguinte árvore de decisão:



Resultados

Com as árvores criadas, podemos então percorrer os dados de teste e verificar quais previsões estariam as corretas. De maneira interessante, as duas árvores chegaram aos mesmos resultados para todos os casos. Assim, as tabelas de confusão ficaram iguais.

Passenger	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Survived	Prediction
1086	2	Drew, Ma	male	8	0	2	28220	32.5		S	1	0
919	3	Daher, Mr	male	22.5	0	0	2698	7.225		C	1	0
1141	3	Khalil, Mr	female		1	0	2660	14.4542		C	0	1
1283	1	Lines, Mrs	female	51	0	1	PC 17592	39.4	D28	S	1	1
1191	3	Johansson	male	29	0	0	347467	7.8542		S	0	0

Tabela de confusão

Confusion Matrix	Positive (Actual)	Negative (Actual)
Positive (Prediction)	True Positive - 1	False Positive - 2
Negative (Prediction)	False Negative - 1	True Negative - 1

As classificações da tabela de confusão foram as seguintes:

True Positive = 1;
True Negative = 1;
False Positive = 2;
False Negative = 1;

Para um total de 2 casos corretos, temos então os cálculos das seguintes métricas:

Precision	TP/(TP+FN) = 2/(2+1) = 0.667
Recall	TP/(TP+FP) = 2/(2+1) =
Error Rate	1-accuracy = 1-0.4 = 0.6
Accuracy	Casos corretos/Casos totais = ⅖ = 0.4

Análise

As duas árvores são bastante diferentes. Enquanto que a árvore gerada pelo ID3 com base na Impureza de Gini era bastante equilibrada, com altura h = 2, enquanto que a árvore resultante do Ganho de Informação tinha uma altura muito superior, com h = 5, sendo que só um existia um ramo principal.

Quanto aos casos de teste, ambas as árvores devolveram os mesmos resultados. Seriam precisos mais casos para além dos atuais para observar melhor a diferença das previsões das duas árvores.

A nível dos dados de treino, talvez uma melhor distribuição da população pelas diferentes classes teria ajudado a árvore do Ganho de Informação a ficar mais equilibrada. Por exemplo, na última faixa etária só existia um elemento. Isso tornava muito fácil dividir a população por esse critério, mas deixava um número grande de pessoas na população resultante. Da mesma forma, valores muito fora da média para idades e preço do bilhete, criaram splits bastante espaçosos, que não representavam bem as populações que neles ficaram inseridos.

Conclusões

Em comparação ao primeiro projeto prático, este foi bastante mais acessível. Embora trabalhoso, a nível da quantidade de cálculos e organização de dados, não terá sido uma tarefa impossível. Algo que não foi um desafio foi a divisão de tarefas, que foram rapidamente distribuídas com o beneplácito do grupo.

O trabalho foi uma grande oportunidade, tanto a nível da aprendizagem do material de estudo abrangido neste projeto, como também a nível de aplicações que não estariam diretamente relacionadas a estes tópicos (como por exemplo, aprender a utilizar o Google Sheets e um mínimo de javascript para a criação de funções personalizadas).