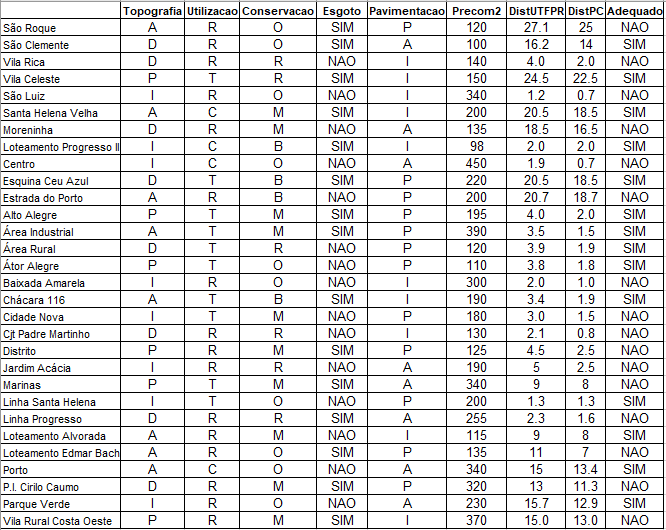
**Base de treino:**



**J48**

O algoritmo **C4.5** constrói árvores de decisão a partir de um conjunto de dados de treinamento da mesma forma que o algoritmo ID3, utilizando o conceito de Entropia. O conjunto de dados de treinamento é um conjunto{\displaystyle S={s\_{1},s\_{2},...}} de amostras já classificadas. Cada amostra {\displaystyle s\_{i}} consiste de um vetor p-dimensional {\displaystyle (x\_{1,i},x\_{2,i},...,x\_{p,i})}, onde o {\displaystyle x\_{j}} representa valores de atributos ou características da amostra, assim como a categoria ou a classe à qual {\displaystyle s\_{i}} pertence.

Em cada nó da árvore, o algoritmo C4.5 escolhe o atributo dos dados que mais efetivamente particiona o seu conjunto de amostras em subconjuntos tendendo a uma categoria ou a outra. O critério de particionamento é o ganho de informação normalizado (diferença em entropia). O atributo com maior ganho de informação normalizado é escolhido para tomar a decisão. O algoritmo C4.5 então repete a etapa anterior nas partições menores.

Este algoritmo possui alguns casos básicos:

* todas as amostras do conjunto pertencem a uma mesma categoria; quando este caso ocorre, o algoritmo simplesmente cria um nó folha para a árvore de decisão e escolhe a categoria em questão;
* nenhuma das características fornece ganho de informação. Neste caso, o algoritmo C4.5 cria um nó de decisão árvore acima usando o valor esperado;
* instâncias previamente não vistas. Novamente, o algoritmo C4.5 cria um nó de decisão árvore acima usando o valor esperado.

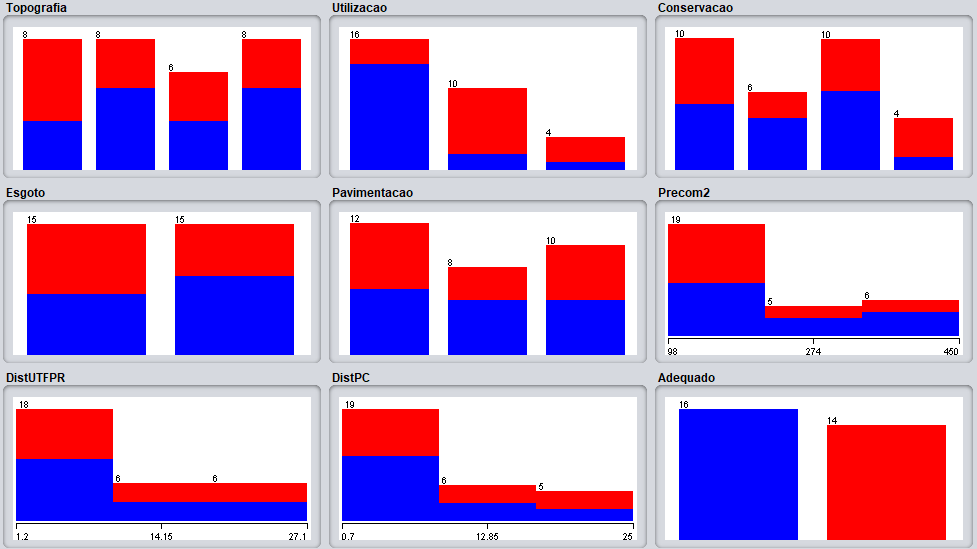
Algoritmo **C4.5**, testado em 7 casos de 10 instâncias cada. As instâncias são derivadas da base de treino (30 instâncias).

A taxa de acerto da base de treino é de 80% com 24 acertos, 6 erros e 0 não classificados.

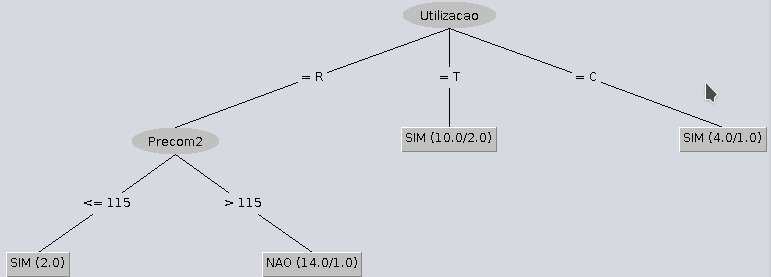
**Matriz de confusão da base:**

|  |  |  |
| --- | --- | --- |
| A | B |  |
| 13 | 3 | A=NÃO |
| 3 | 11 | B=SIM |

Por 3 vezes o classificador inferiu que a resposta era NÃO e na verdade era SIM. E por 3 vezes o classificador inferiu que a resposta era SIM e na verdade era NÃO.

**Gráfico dos atributos:** 

**Árvore:**



Com base na árvore definida e no gráfico dos atributos, podemos analisar que o atributo **Precom2** mescla discretamente as saídas de **SIM** e **NÃO**, permitindo uma faixa de erro.

**Modificações:**

**Teste 01:** -

**Teste 02:**

Linha 4 - Precom2 -> de 140 para 290

Linha 8 - Precom2 -> de 135 para 200

**Teste 03:**

Linha 5 - DistUTFPR -> de 150 para 30

Linha 9 - DistUTFPR -> de 98 para 290

**Teste 04:**

Linha 7 - DistPC -> de 18.5 para 14.3

Linha 8 - DistPC -> de 16.5 para 3.0

**Teste 05:**

Linha 2 - Precom (de 120 para 80) - DistUTFPR (de 27.1 para 18.0) - DistPC (de 25 para 10)

Linha 4 - Precom (de 140 para 100) - DistUTFPR ( de 4.0 para 12.5) - DistPC (de 2.0 para 9.0)

Linha 5 - Precom (de 150 para 100)

Linha 8 - DistUTFPR (de 18.5 para 9.6)

Linha 9 - DistPC de 2.0 para 25.0)

**Teste 06:**

Linha 4 - Conservacao ( de R para O) - Pavimentacao ( I para A )

Linha 7 - Conservacao ( de M para O) - Adequado ( de SIM para NAO )

Linha 11 - Topografia (de D para A) - Adequado (de SIM para NAO) - Pavimentacao ( de P para A )

**Teste 07:**

Linha 5 - Utilizacao (de T para C) - Precom2 (de 150 para 200)

Linha 6 - Utilizacao (de R para T) - Precom2 (de 340 para 120)

Linha 7 - Utilizacao (de C para R) - Precom2 (de 200 para 360)

**Matriz de confusão dos testes:**

**Teste 01:**

|  |  |  |
| --- | --- | --- |
| A | B |  |
| 4 | 1 | A=NÃO |
| 0 | 5 | B=SIM |

**Teste 02:**

|  |  |  |
| --- | --- | --- |
| A | B |  |
| 4 | 1 | A=NÃO |
| 0 | 5 | B=SIM |

**Teste 03:**

|  |  |  |
| --- | --- | --- |
| A | B |  |
| 4 | 1 | A=NÃO |
| 0 | 5 | B=SIM |

**Teste 04:**

|  |  |  |
| --- | --- | --- |
| A | B |  |
| 4 | 1 | A=NÃO |
| 0 | 5 | B=SIM |

**Teste 05:**

|  |  |  |
| --- | --- | --- |
| A | B |  |
| 2 | 3 | A=NÃO |
| 0 | 5 | B=SIM |

**Teste 06:**

|  |  |  |
| --- | --- | --- |
| A | B |  |
| 4 | 3 | A=NÃO |
| 0 | 3 | B=SIM |

**Teste 07:**

|  |  |  |
| --- | --- | --- |
| A | B |  |
| 4 | 1 | A=NÃO |
| 0 | 5 | B=SIM |

**Tabela de testes:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| TESTE | ACERTOS | ERROS | %ACERTOS | %SIM | %NÃO |
| 1 | 9 | 1 | 90 | 83,3 | 100 |
| 2 | 9 | 1 | 90 | 83,3 | 100 |
| 3 | 9 | 1 | 90 | 83,3 | 100 |
| 4 | 9 | 1 | 90 | 83,3 | 100 |
| 5 | 7 | 3 | 70 | 62,5 | 100 |
| 6 | 7 | 3 | 70 | 50 | 100 |
| 7 | 9 | 1 | 90 | 83,3 | 100 |
| **TOTAIS** | **8,428571** | **1,571429** | **84,28571** | **75,57143** | **100** |

Para os casos de teste acima mostrados, o algoritmo C4,5 teve uma taxa de 84,2% de acertos. Nos casos em que a resposta foi **NÃO**, o algoritmo acertou 100%.

**ID3**

O algoritmo ID3 começa com o conjunto original {\ displaystyle S}como o nó raiz. Em cada iteração do algoritmo, itera através de cada atributo não utilizado do conjunto {\ displaystyle S}e calcula a entropia {\ displaystyle \ mathrm {H} {(S)}}(ou ganho de informação {\ displaystyle IG (S)}) desse atributo. Em seguida, ele seleciona o atributo que possui o

menor valor de entropia (ou maior ganho de informações). O conjunto {\ displaystyle S}é então dividido ou particionado pelo atributo selecionado para produzir subconjuntos dos dados. (Por exemplo, um nó pode ser dividido em nós filhos com base nos subconjuntos da população cujas idades são menores que 50, entre 50 e 100 e maiores que 100.) O algoritmo continua a se repetir em cada subconjunto, considerando apenas atributos nunca selecionado antes.

A recursão em um subconjunto pode parar em um desses casos:

* cada elemento no subconjunto pertence à mesma classe. Nesse caso, o nó é transformado em um nó de folha e rotulado com a classe dos exemplos.
* não há mais atributos a serem selecionados, mas os exemplos ainda não pertencem à mesma classe. Nesse caso, o nó é formado por um nó folha e rotulado com a classe mais comum dos exemplos no subconjunto.
* não há exemplos no subconjunto, o que acontece quando nenhum exemplo no conjunto pai foi encontrado para corresponder a um valor específico do atributo selecionado. Um exemplo pode ser a ausência de uma pessoa entre a população com mais de 100 anos. Em seguida, um nó folha é criado e rotulado com a classe mais comum dos exemplos no conjunto do nó pai.

Ao longo do algoritmo, a árvore de decisão é construída com cada nó não terminal (nó interno) representando o atributo selecionado no qual os dados foram divididos e os nós terminais (nó folha) representando o rótulo de classe do subconjunto final dessa ramificação.

Algoritmo **ID3**, testado em 3 casos de 10 instâncias cada. As instâncias são derivadas da base de treino (30 instâncias).

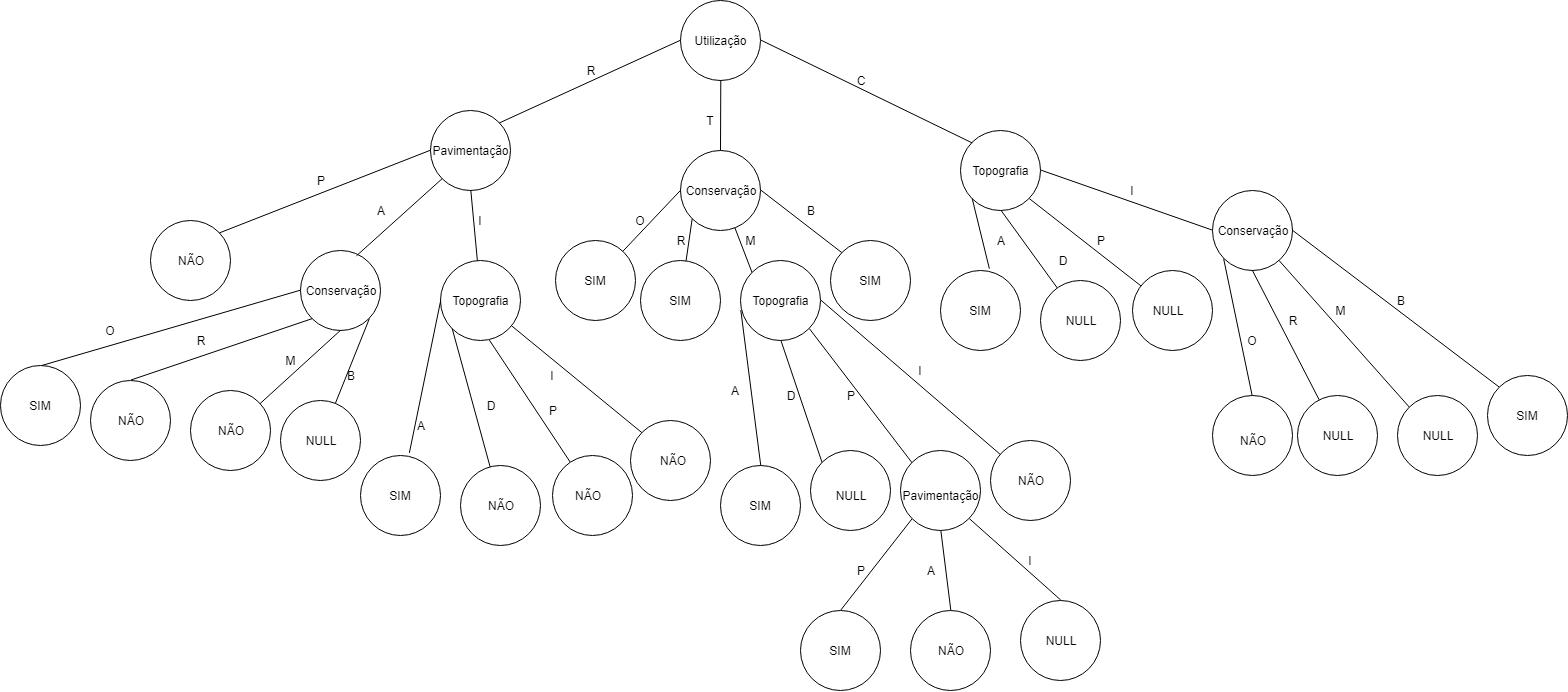
A taxa de acerto da base de treino é de 63,3% com 19 acertos, 8 erros e 3 não classificados.

**Matriz de confusão da base:**

|  |  |  |
| --- | --- | --- |
| A | B |  |
| 11 | 3 | A=NÃO |
| 5 | 8 | B=SIM |

Por 3 vezes o classificador inferiu que a resposta era SIM e na verdade era NÃO. E por 5 vezes o classificador inferiu que a resposta era NÃO e na verdade era SIM.

**Árvore:**



**Tabela de testes:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| TESTE | ACERTOS | ERROS | %ACERTOS | %SIM | %NÃO |
| 1 | 10 | 0 | 100 | 100 | 100 |
| 2 | 10 | 0 | 100 | 100 | 100 |
| 3 | 10 | 0 | 100 | 100 | 100 |
| **TOTAIS** | **10** | **0** | **100** | **100** | **100** |

**Modificações:**

**Teste 01:**

Linha 3 - Topografia (de D para A)

Linha 6 - Topografia (I para D)

Linha 11 - Topografia (D para I)

**Teste 02:**

Linha 4 - Conservacao (de R para O)

Linha 6 - Conservacao (de O para M)

Linha 9 - Conservacao (de B para M)

**Teste 03:**

Linha 5 - Esgoto (de SIM para NAO) - Pavimentacao (de I para P)

Linha 9 - Esgosto (de SIM para NAO) - Pavimentacao (de I para A)

Linha 10 - Esgosto (de NAO para SIM) - Pavimentacao (de A para I)

**Matriz de confusão dos testes:**

**Teste 01:**

|  |  |  |
| --- | --- | --- |
| A | B |  |
| 5 | 0 | A=NÃO |
| 0 | 5 | B=SIM |

**Teste 02:**

|  |  |  |
| --- | --- | --- |
| A | B |  |
| 5 | 0 | A=NÃO |
| 0 | 5 | B=SIM |

**Teste 03:**

|  |  |  |
| --- | --- | --- |
| A | B |  |
| 5 | 0 | A=NÃO |
| 0 | 5 | B=SIM |

Para os casos de teste acima mostrados, o algoritmo ID3 teve uma taxa de 100% de acertos.

**Resultados:**

Todos os testes foram feitos sobre uma base de 30 instâncias.

Para os casos testados, apesar de o algoritmo ID3 possuir grandes chances de não classificar algumas amostras, seu desempenho foi superior obtendo 100% de acerto em 3 instâncias, para classificar sobre a base, seu tempo de execução foi de 0,02s.

Sobre o algoritmo C4.5, foram classificadas 7 instâncias, seu tempo de execução para classificar a base foi de 0s, porém o desempenho foi menor se comparado ao ID3, a taxa de acerto foi de 84,2%. Todas as vezes em que a resposta é **NÃO**, a classificação foi feita de maneira correta, os falsos positivos ocorrem quando a resposta final deve ser **SIM**.