Problema2 – Evasão de alunos da UFCG dos cursos de Engenharia Mecânica e Ciência da Computação.

Primeiro decidi criar uma arvore com todas as variáveis para ser possível que a arvore de decisão me mostrasse quais são as mais importantes e menos.

```
Evaluation on training data (2109 cases):

     Decision Tree
   ----------------
   Size      Errors

     27   72( 3.4%)   <<


    (a)   (b)    <-classified as
   ----  ----
   1905    10    (a): class 0
     62   132    (b): class 1


 Attribute usage:

 100.00% SITUACAO
  88.29% MEDIA
  81.08% CODIGO
  15.36% MATRICULA
  14.37% PERIODO
  10.76% CURSO
   8.16% DEPARTAMENTO
```

Vendo os dados acima, temos três atributos muito usados pela arvore de decisão: Situacao, media e código. Além de um erro de apenas 3.4%. É estranho que o departamento da disciplina cursada tenha tão pouco envolvido na evasão do aluno.

Mas tendo em vista esses resultados, partimos para geração da arvore so com os atributos mais significativos, para verificar se existe alguma melhora.

```
Evaluation on training data (2109 cases):

     Decision Tree
   ----------------
   Size      Errors

      4  166( 7.9%)   <<


    (a)   (b)    <-classified as
   ----  ----
   1890    25    (a): class 0
    141    53    (b): class 1


 Attribute usage:

 100.00% dados.SITUACAO
  11.71% dados.CODIGO
```

Podemos ver que o erro cresceu e que o atributo media, estranhamente, não foi mais relevante para a arvore.

Tendo em vista esses resultados, vamos partir para a criação do modelo para os cursos individuais e ver se as métricas são similares e se os atributos a serem usados devem ser os mesmos.

Abaixo temos os dados da arvore para Ciência da Computação:

```
Evaluation on training data (1304 cases):

     Decision Tree
   ----------------
   Size      Errors

     18   49( 3.8%)   <<


    (a)   (b)    <-classified as
   ----  ----
   1153     7    (a): class 0
     42   102    (b): class 1


 Attribute usage:

 100.00% SITUACAO
  90.64% MEDIA
  78.99% CODIGO
  16.41% MATRICULA
  11.66% DEPARTAMENTO
  11.12% PERIODO
```

Podemos ver que Media toma uma porcentagem mais significativa, embora o erro permaneça próximo ao exemplo do modelo geral visto acima. Outra mudança menos significativa é que Departamento passa a ser mais significativo também.

Tendo em vista esse resultado, vamos ver como ficam os resultados da arvore para os três atributos mais significativos:

```
Evaluation on training data (1304 cases):

     Decision Tree
   ----------------
   Size      Errors

      2  112( 8.6%)   <<


    (a)   (b)    <-classified as
   ----  ----
   1115    45    (a): class 0
     67    77    (b): class 1


 Attribute usage:

 100.00% dados.SITUACAO
```

Temos um exemplo parecido, novamente, com o caso equivalente do modelo geral. O erro também é um pouco maior que o modelo com todos os atributos. Mas nesse caso especifico é (do modelo sem atributos) a arvore ficou formada apenas por um atributo (a Situação do aluno), o que não me parece uma boa previsão por si só.

Continuando com os modelos, vamos agora olhar o modelo para Engenharia mecânica:

```
Evaluation on training data (1304 cases):

     Decision Tree
   ----------------
   Size      Errors

     18   49( 3.8%)   <<


    (a)   (b)    <-classified as
   ----  ----
   1153     7    (a): class 0
     42   102    (b): class 1


 Attribute usage:

 100.00% SITUACAO
  90.64% MEDIA
  78.99% CODIGO
  16.41% MATRICULA
  11.66% DEPARTAMENTO
  11.12% PERIODO
```

Podemos ver que o exemplo é muito próximo ao visto no modelo de Computação, com porcentagens bem parecidas também.

Tendo em vista esse resultado, vamos ver como ficam os resultados da arvore para os três atributos mais significativos:

```
Evaluation on training data (1304 cases):

     Decision Tree
   ----------------
   Size      Errors

      2  112( 8.6%)   <<


    (a)   (b)    <-classified as
   ----  ----
   1115    45    (a): class 0
     67    77    (b): class 1


 Attribute usage:

 100.00% dados.SITUACAO
```

Temos um exemplo parecido com os modelos vistos a cima. E como no caso do modelo de computação a arvore também ficou formada apenas por um atributo (a Situação do aluno), o que não me parece uma boa previsão por si só.

Tendo em vista essas analises, deveremos usar os modelos com todos os atributos para fazer as previsões dos arquivos de teste. Não só pelo erro indicado pela ferramenta ser menor, mas também pelo tamanho da arvore gerada, que passa uma segurança um pouco maior (afinal uma arvore de altura um pode não ser muito significativo)