RELATÓRIO DO EXERCÍCIO DE FIXAÇÃO DE CONCEITOS – EFC 2

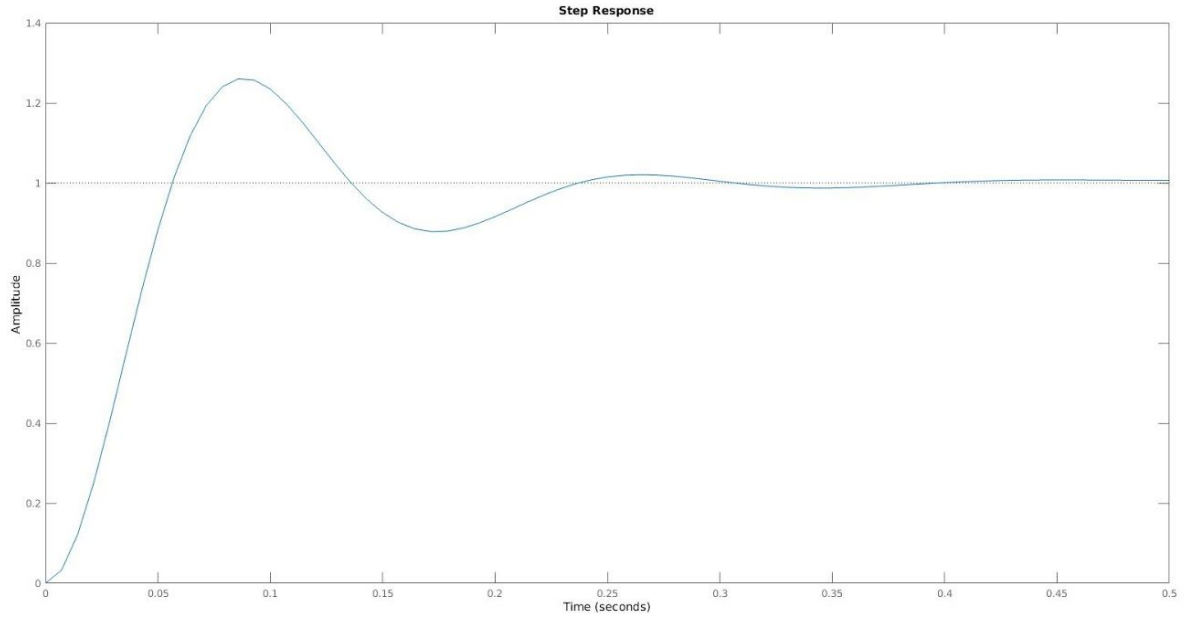
EA072 - Inteligência Artificial em Aplicações Industriais - Turma A

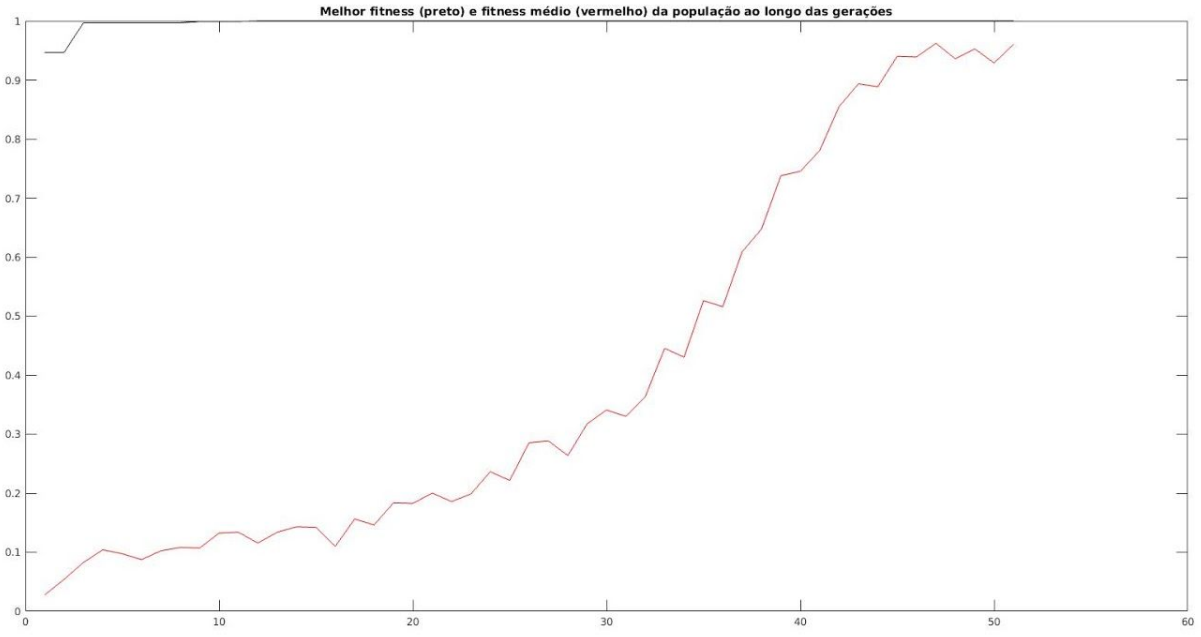
Aluno: William Quintas de Melo - RA: 188684 / **Professor:** Fernando J. Von Zuben

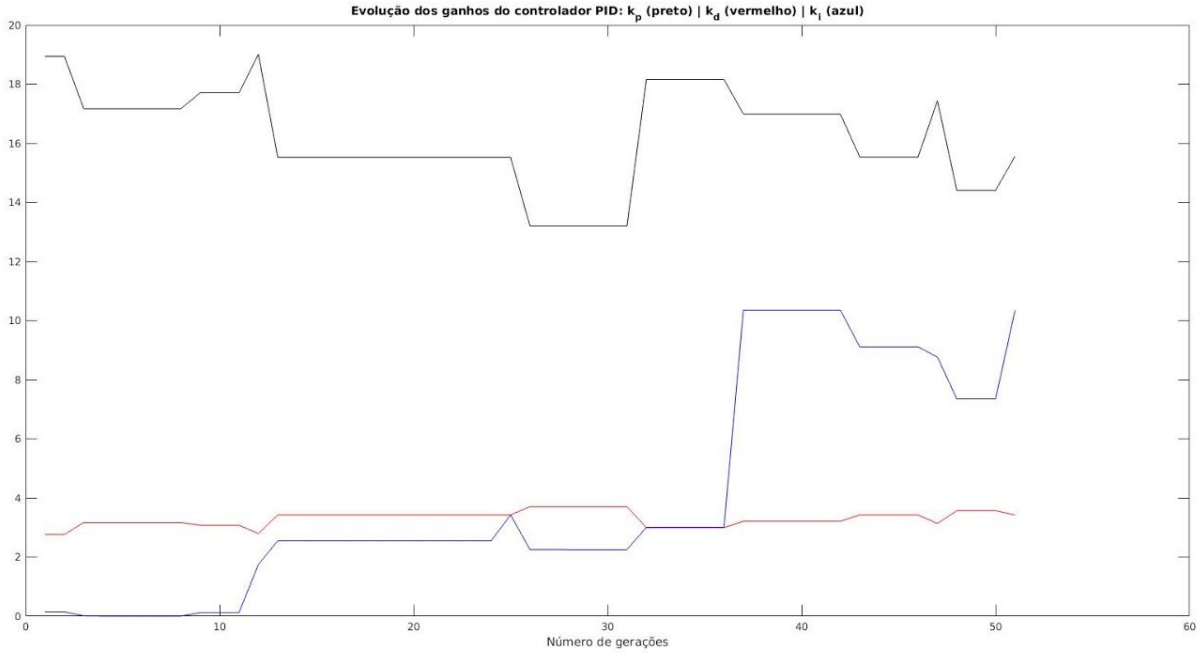
QUESTÃO 1

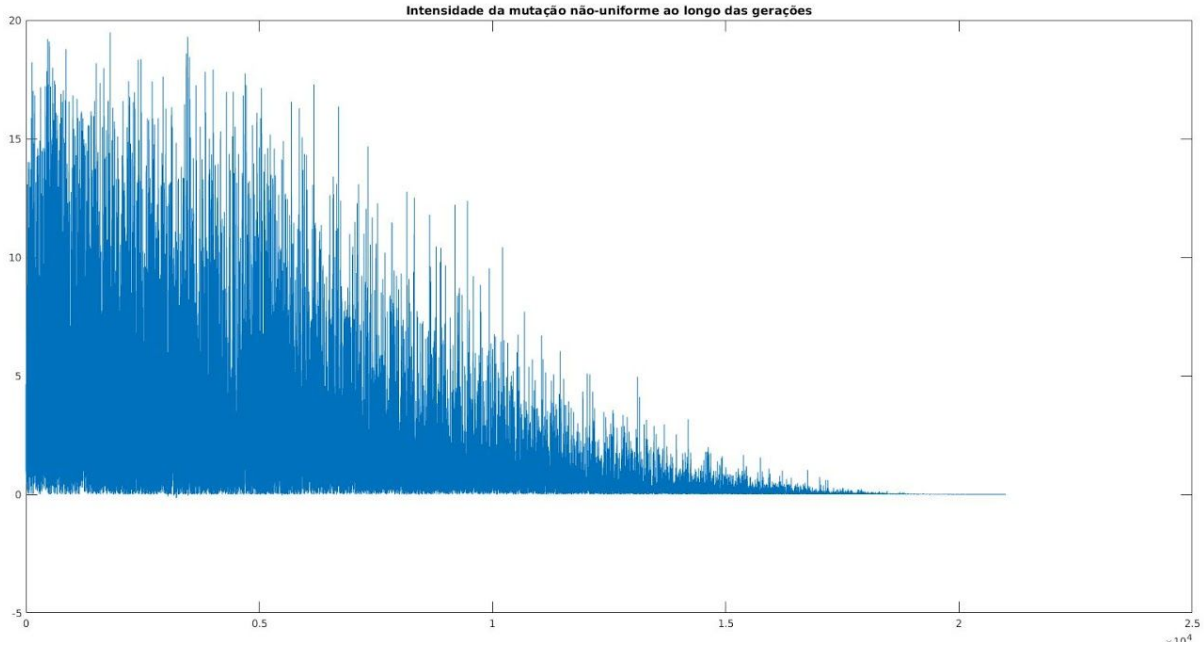
Para a realização desta questão, foi executado no MATLAB o programa **prog_Q1_EFC2.m** fornecido pelo professor no toolbox. Já na primeira execução pôde-se obter os 4 gráficos abaixo:

Resposta ao degrau









Após a obtenção dos gráficos, estes foram analisados em busca de informações acerca da solução obtida para o problema.

O primeiro gráfico retrata a resposta ao degrau da última geração, que nos dá os valores dos ganhos para a solução ótima. Combinando este gráfico com conhecimentos de controle, inclusive do enunciado, nota-se como a resposta é rápida. O gráfico do melhor fitness e do fitness médio mostra que o valor do melhor fitness atingiu 1 em pouco tempo. Outro ponto importante de se notar desse gráfico é que o fitness médio apresenta uma evolução crescente durante a execução, sendo que ao final este valor está entre 0,9 e 1. O gráfico da evolução dos ganhos apresenta a maneira como o melhoramento dos valores dos ganhos avança durante a execução do programa. Em conjunto com o gráfico do melhor fitness e do fitness

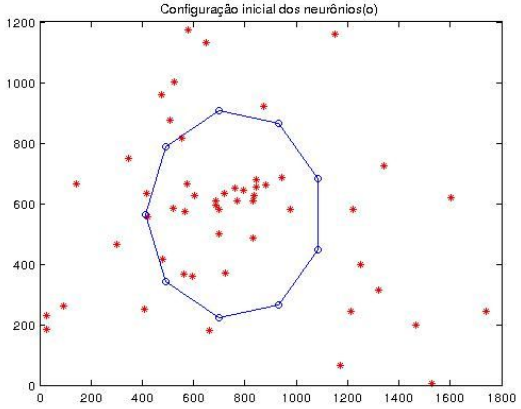
médio, pode se analisar como os valores dos ganhos influenciam na melhora do fitness, sendo possível obter os valores que fornecem a solução ótima para o fitness. Já conjuntamente com o gráfico da resposta ao degrau, pode se explicar a rapidez da resposta ao impulso, pois já que os valores dos ganhos são ótimos, espera-se que o sistema se controle de forma também rápida. Por fim, o gráfico da intensidade de mutação exibe a queda no número de mutações ao longo da execução. Esse comportamento pode ser explicado pela maior proximidade das gerações mais recentes com a solução ótima, o que faz com que não haja um índice de mutação muito elevado.

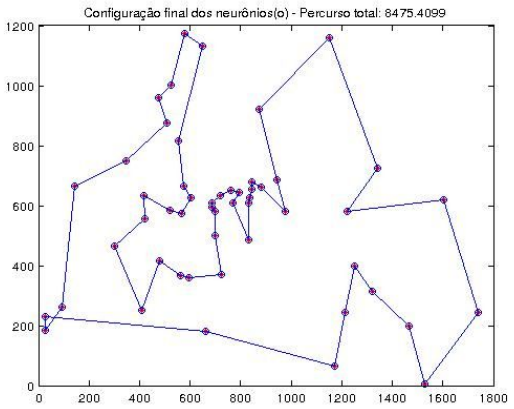
QUESTÃO 3

Foi obtida a solução de uma instância com pelo menos 50 cidades do problema de caixeiro viajante por mapa de Kohonen através de 3 execuções do toolbox fornecido de autoria do Prof. Levy Boccato (FEEC/Unicamp) para cada arquivo de apoio fornecido:

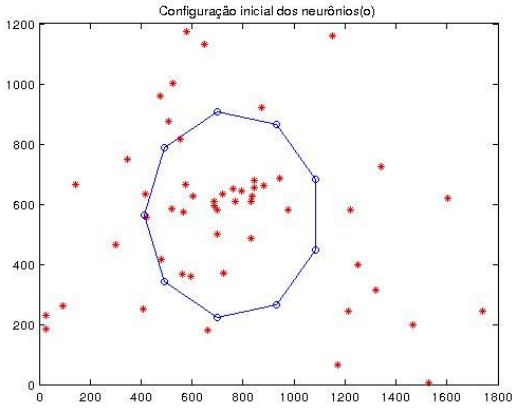
Berlin52

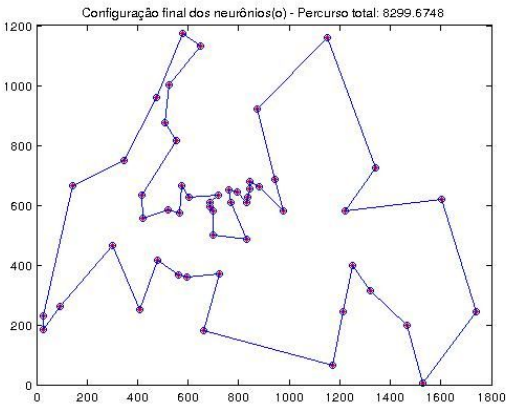
Primeira execução:



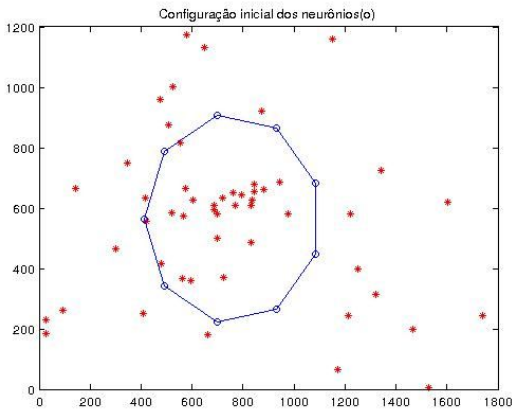


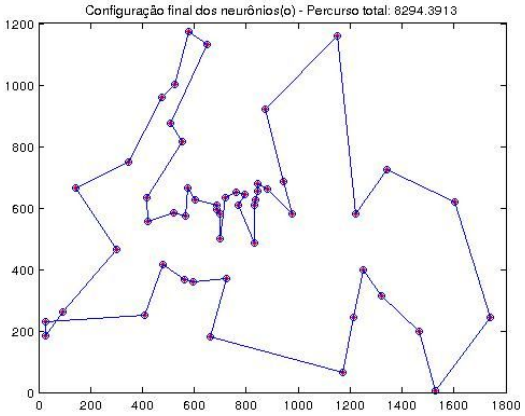
Segunda execução:





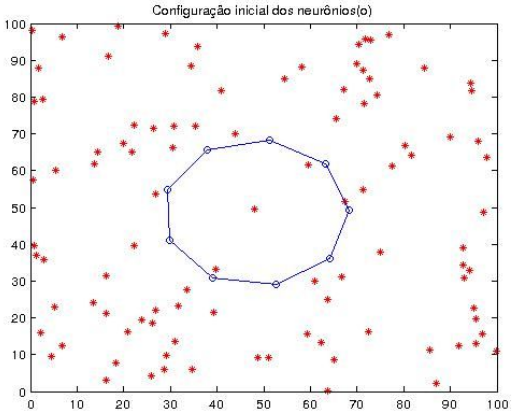
Terceira execução:

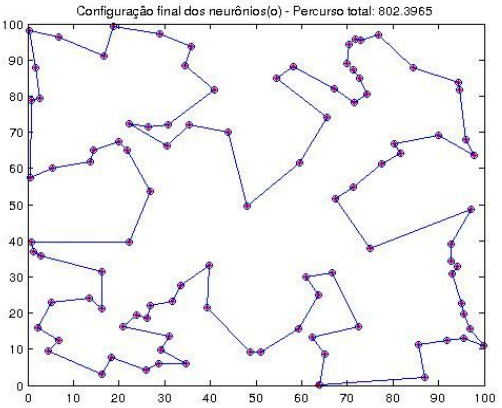




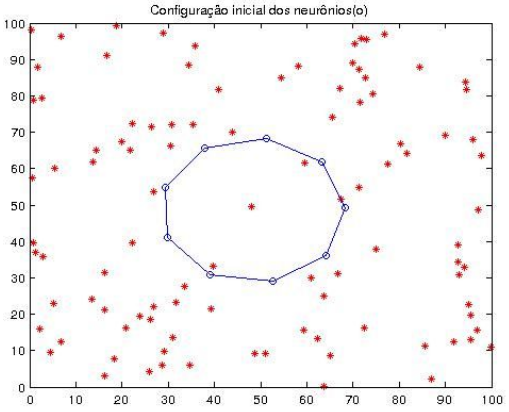
Inst1

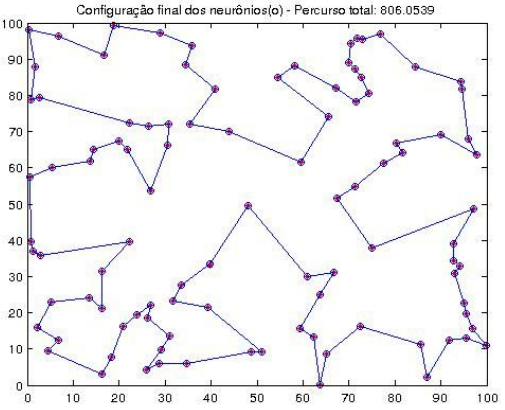
Primeira execução:



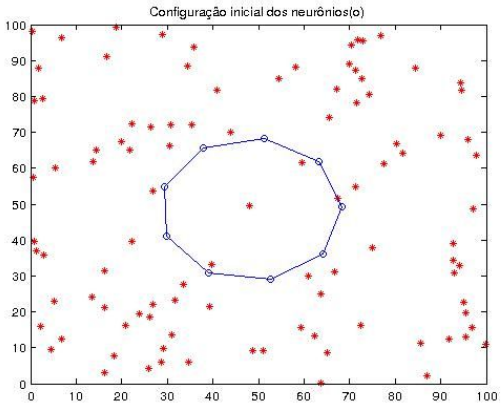


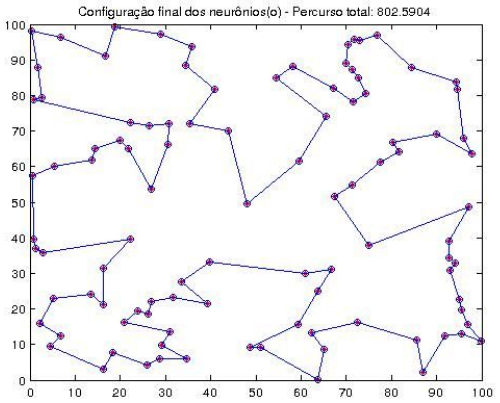
Segunda execução:





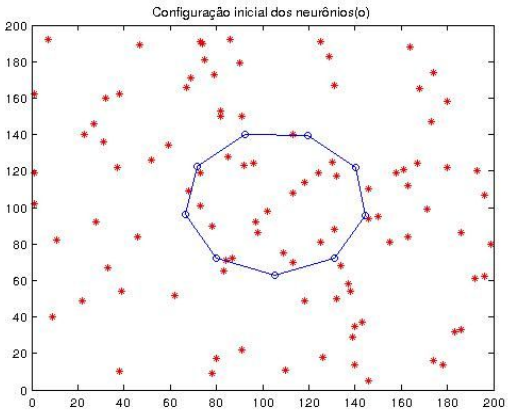
Terceira execução:

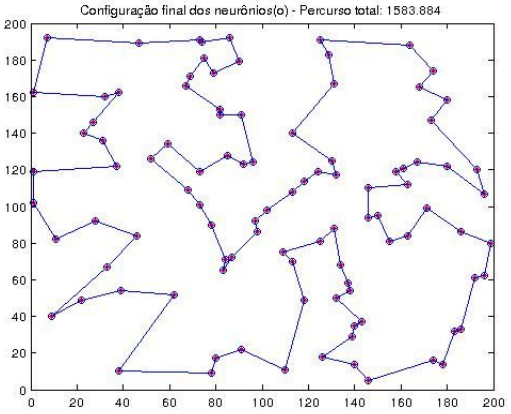




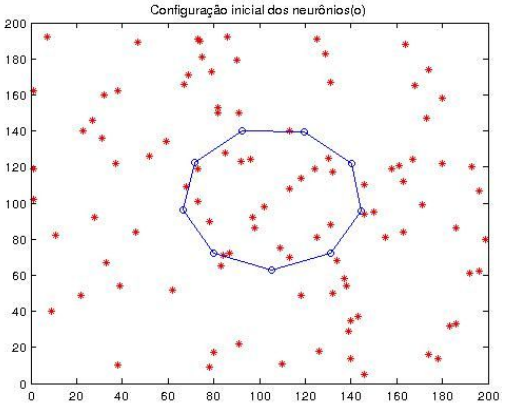
Inst2

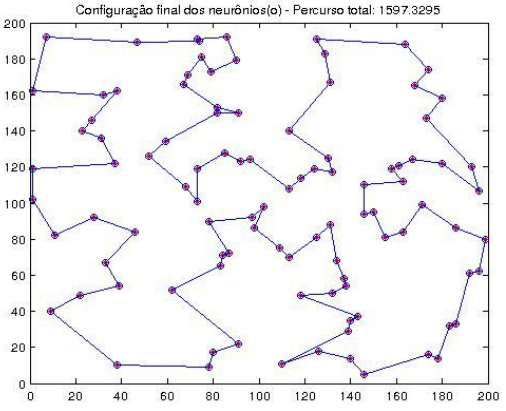
Primeira execução:



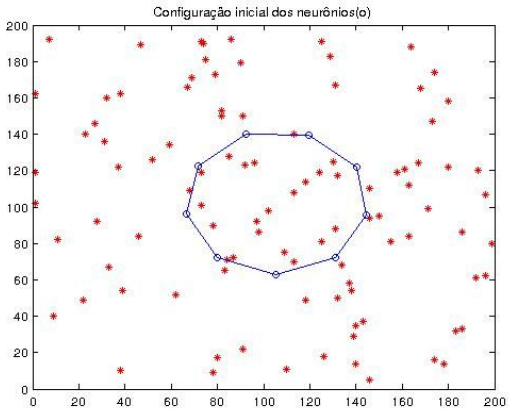


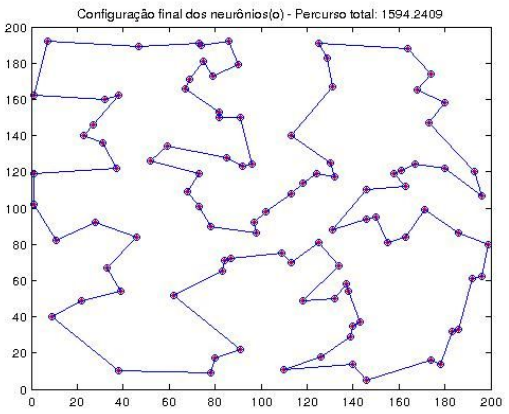
Segunda execução:





Terceira execução:





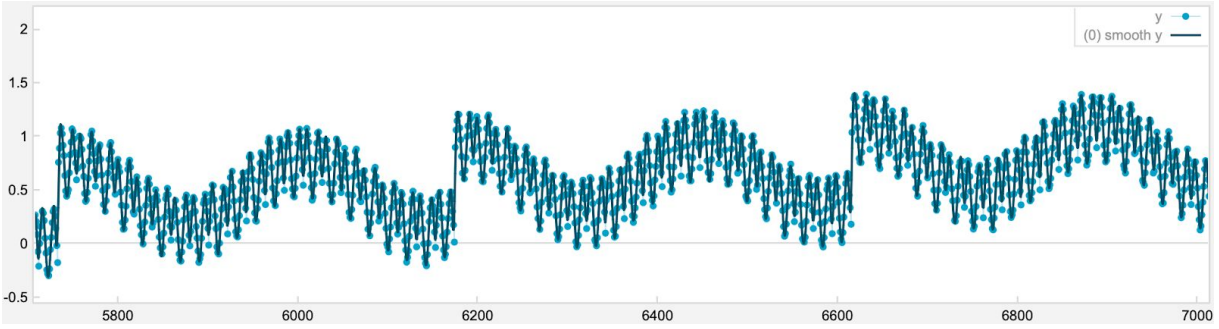
QUESTÃO 4

Para um primeiro contato com o Eureqa, foi feito o experimento de gerar um mapeamento $\mathbb{R}^1 \rightarrow \mathbb{R}^1$ escolhendo a função $y(x) = \frac{1}{3}cos(x)sen(x-3)$ que pode ser reescrita como $y(x) = \frac{1}{6}(-sen(3-2x) - sen(3))$ para um intervalo de 0 a 10 com passo de 0,1 e adicionando um ruído ao y(x) calculado. Depois, se empregou o software Eureqa para propor alternativas de funções de aproximação. A solução escolhida com boa relação de custo-benefício entre acurácia e simplicidade foi $y(x) = 0,024357279 + 0,169826059990119sen(3,35318784894538 + 1,98086456770817x)$. Abaixo são exibidas informações acerca dessa solução, obtidas na interface do Eureqa:

- Qualidade do ajuste: 0,9868186
- Coeficiente de correlação: 0,9946428
- Erro máximo: 0,04156145
- Erro quadrático médio: 0,00018993122
- Erro absoluto médio: 0,010921259
- Coeficientes: 3
- Complexidade: 10
- Objetivo primário: 0,089477554
- Fit (Objetivo primário normalizado): 0,1020531

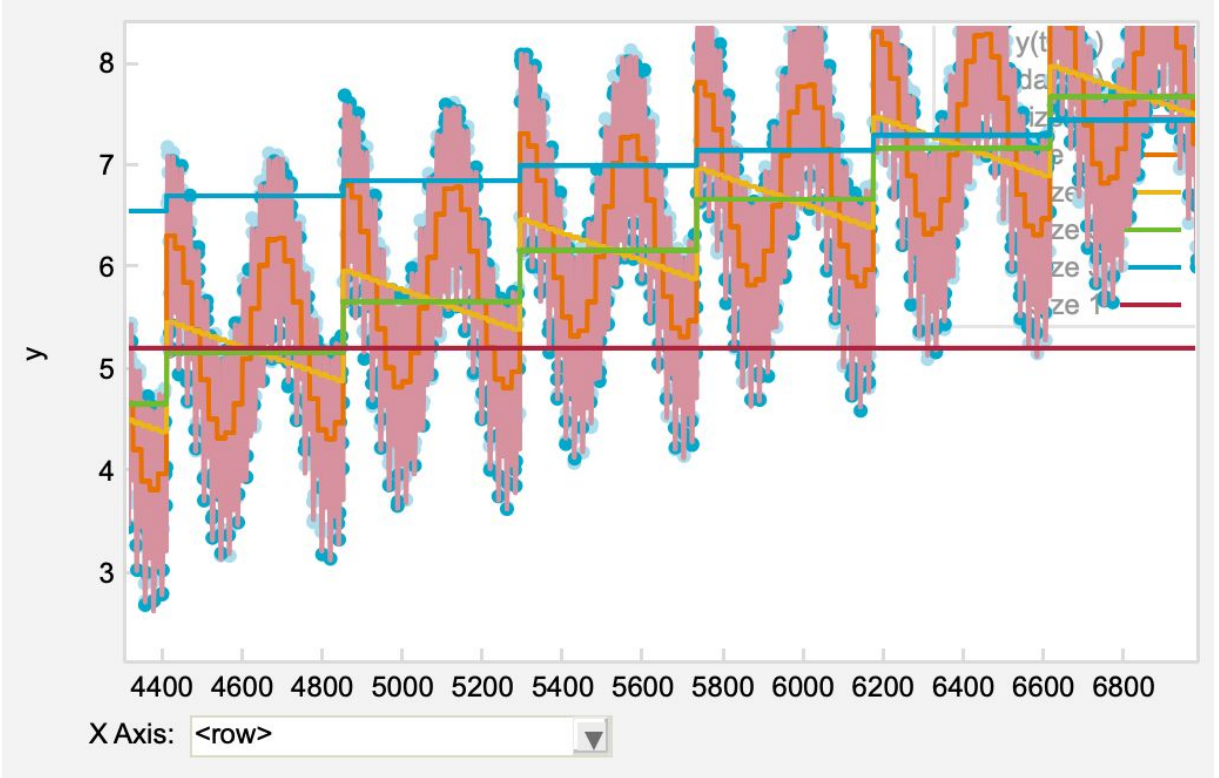
Depois disso, assim como no mapeamento $\mathbb{R}^1 \rightarrow \mathbb{R}^1$, foi escrito um programa em MATLAB para um mapeamento $\mathbb{R}^3 \rightarrow \mathbb{R}^1$, ou seja, $y = f(x_1, x_2, x_3)$. A função escolhida foi $y(x) = x_1 + cos(x_2) + sen(x_3)$ para um intervalo de 0 a 10 em cada variável com passo de 0,5 e adicionando um ruído ao y calculado. Depois, se empregou novamente o software Eureqa, preenchendo os campos de dados com os gerados pelo programa do MATLAB.

O gráfico abaixo mostra um trecho do gráfico de y(n) - onde n é o número do conjunto de dados (x_1, x_2, x_3):



Foi executada então a busca no software Eureqa, obtendo algumas soluções que são mostradas abaixo:

Tamanho	Fit	Solução
19	0,018	y = 0.0972898639474247 + 1.00031577971066*x1 + 1.00031577971066*cos(x2) + 0.995783589701748*sin(x3)
12	0,210	y = 0.305995115795059 + 1.00119319868786*x1 + 0.998172570804925*cos(x2)
9	0,287	y = 0.438359566548789 + 1.00423282971455*x1 - 0.0571342057057035*x2
5	0,293	y = 0.136804309746085 + 1.00480907203521*x1
3	0,839	y = 5.21150507841054 + 0.297969625552048*x1
1	1,000	y = 5.20651885005326

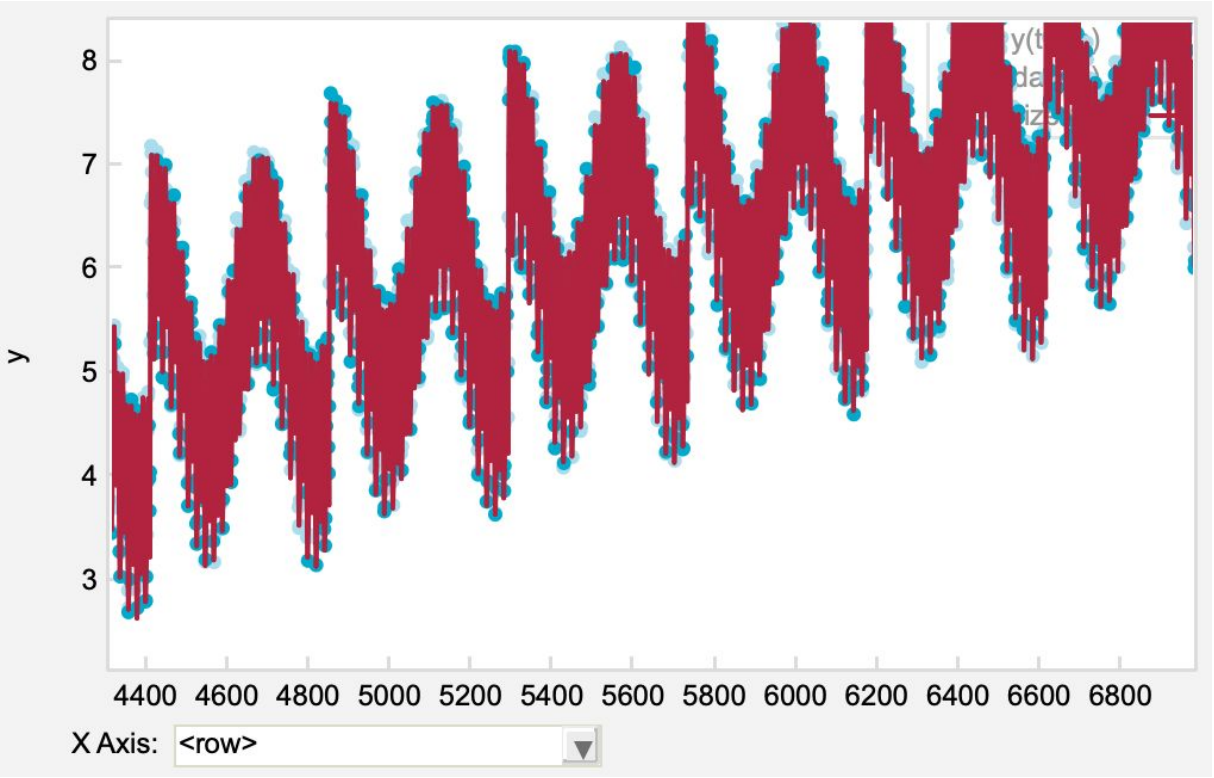


A solução escolhida com boa relação de custo-benefício entre acurácia e simplicidade foi $y(x_1, x_2, x_3) = 0,097289864 + 1,000315780x_1 + 1,000315780cos(x_2) + 0,995783590sin(x3)$

. Abaixo são exibidas informações acerca dessa solução, obtidas na interface do Eureqa:

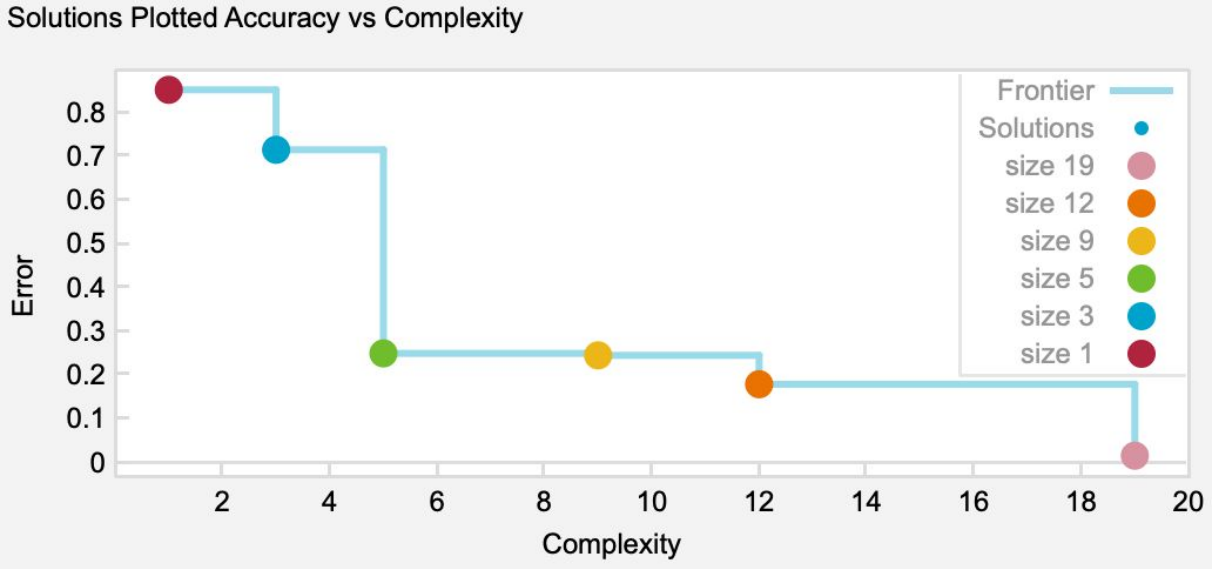
- Qualidade do ajuste: 0,99927313
- Coeficiente de correlação: 0,99963703
- Erro máximo: 0,97498343
- Erro quadrático médio: 0,0073361086
- Erro absoluto médio: 0,048824965
- Coeficientes: 4
- Complexidade: 19
- Objetivo primário: 0,015327717
- Fit (Objetivo primário normalizado): 0,017997705

Abaixo há uma imagem com o gráfico dessa solução y(n):



Nota-se que, dada a devida escala, ela se parece bem com o gráfico inserido dos dados, e até a equação ficou com forma bem similar à função utilizada para gerar os dados.

A imagem abaixo exibe a fronteira de Pareto obtida para o mapeamento $\mathbb{R}^3 \rightarrow \mathbb{R}^1$. A fronteira de pareto é formada por pontos que indicam a escolha ótima em termos de eficiência dado o conjunto de opções.



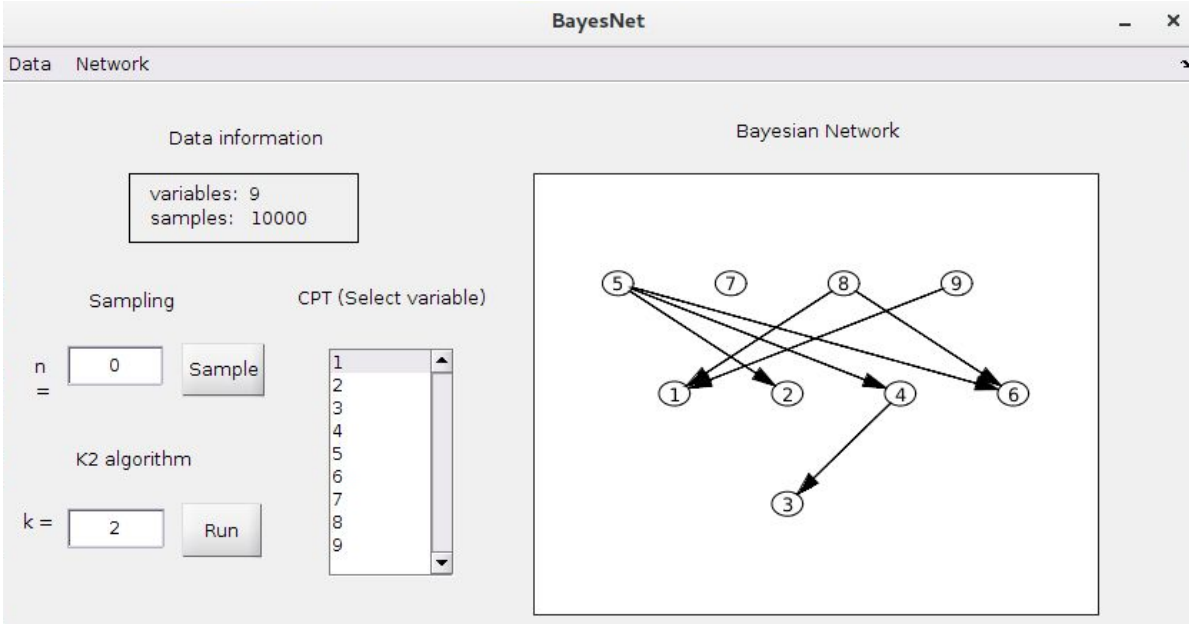
QUESTÃO 5

(a) Apresente as características dos dados fornecidos: número de amostras, número de atributos, valores assumidos pelos atributos (no caso de atributos binários, 1 é falso e 2 é verdadeiro);

Os dados fornecidos possuem 10000 amostras, com 9 atributos, sendo que todos os atributos, com exceção de um, são atributos binários. O atributo da segunda coluna dos dados apresenta valores inteiros de 1 a 7.

(b) Apresente a rede bayesiana resultante, incluindo as tabelas de probabilidades associadas a cada nó da rede.

A rede Bayesiana resultante é exibida na imagem abaixo:



As tabelas de probabilidades associadas a cada nó são exibidas abaixo:

VARIÁVEL 1		Falso	Verdadeiro
var. 8 falso	var. 9 falso	1	0
var. 8 falso	var. 9 verdadeiro	0	1
var. 8 verdadeiro	var. 9 falso	0	1
var. 8 verdadeiro	var. 9 verdadeiro	1	0

VARIÁVEL 2	1	2	3	4	5	6	7
var. 5 falso	0.2633	0.1904	0.1769	0.1641	0.1182	0.0655	0.0216
var. 5 verdadeiro	0.1159	0.1414	0.1283	0.1261	0.1657	0.1548	0.1677

VARIÁVEL 3	falso	verdadeiro
var. 4 falso	0.3373	0.6627
var. 4 verdadeiro	0.9802	0.0198

VARIÁVEL 4	falso	verdadeiro
var. 5 falso	0.5949	0.4051
var. 5 verdadeiro	0.7509	0.2491

VARIÁVEL 5	falso	verdadeiro
	0.1481	0.8519

VARIÁVEL 6		Falso	Verdadeiro
var. 5 falso	var. 8 falso	0.8788	0.1212
var. 5 falso	var. 8 verdadeiro	0.5888	0.4112
var. 5 verdadeiro	var. 8 falso	0.0528	0.9472
var. 5 verdadeiro	var. 8 verdadeiro	0.1587	0.8413

VARIÁVEL 7	falso	verdadeiro
	0.5028	0.4972

VARIÁVEL 8	falso	verdadeiro
	0.4820	0.5180

VARIÁVEL 9	falso	verdadeiro
	0.5022	0.4978

(c) Procure justificar por que o atributo 7 não possui conexões a outros nós.

O fato do atributo 7 não possuir conexões a outros nós ocorre porque durante a inferência no processo de construção da rede Bayesiana entendeu-se que este atributo não influencia e nem é influenciado pela probabilidade de ocorrência de nenhuma outra.

(d) Qual é a probabilidade do atributo 5 ser verdade?

Da tabela da variável 5, a probabilidade do atributo 5 ser verdade é 0,8519.

(e) Qual é a probabilidade do atributo 6 ser verdade dado que 5 é falso e 8 é verdade?

Da tabela da variável 6, a probabilidade do atributo 6 ser verdade dado que 5 é falso e 8 é verdade é 0,4112.

(f) Observando a tabela de probabilidades apresentada pelo atributo 1, qual é a relação lógica que você supõe existir entre ele e os atributos 8 e 9?

Ao analisar a tabela da variável 1, supõe-se que há uma relação de XOR entre os atributos 8 e 9.

(g) Qual é a probabilidade do atributo 5 ser verdade, dado que 3 é verdade? Observação: Aqui são necessários cálculos a partir das tabelas de probabilidades, enquanto os itens (d) e (e) saem direto de campos dessas tabelas.

A probabilidade do atributo 5 ser verdade, dado que 3 é verdade, pode ser calculada utilizando o teorema de Bayes, chegando ao resultado de que $P(5 = 1 \mid 3 = 1) = 87.38\%$.

QUESTÃO 6

Dentro do Weka Explorer foi aberto o arquivo fornecido **dados_AD_EFC2.arff** que contém dados reais vinculados à presença ou não de um espécime em certos locais de observação, sob certas condições ambientais.

Foram observados os atributos e estatísticas acerca de seus intervalos de excursão sendo notados 7 atributos, sendo 5 numéricos (temperatura média, umidade média, altura da chuva mensal, precipitação em 21 dias e número de dias de chuva) e 2 categóricos (local e classe). As estatísticas referentes a esses atributos são apresentadas abaixo:

Name: temperatura_media Type: Numeric
Missing: 0 (0%) Distinct: 14 Unique: 0 (0%)

Statistic	Value
Minimum	12.82
Maximum	21.98
Mean	18.448
StdDev	2.965

Name: umidade_media Type: Numeric
Missing: 0 (0%) Distinct: 14 Unique: 0 (0%)

Statistic	Value
Minimum	68.38
Maximum	97.83
Mean	90.306
StdDev	7.247

Name: altura_chuva_mensal Type: Numeric
Missing: 0 (0%) Distinct: 14 Unique: 0 (0%)

Statistic	Value
Minimum	9
Maximum	273.7
Mean	107.779
StdDev	71.769

Name: precip_21_dias Type: Numeric
Missing: 0 (0%) Distinct: 14 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	216.9
Mean	85.179
StdDev	62.745

Name: numero_dias_chuva Type: Numeric
Missing: 0 (0%) Distinct: 11 Unique: 0 (0%)

Statistic	Value
Minimum	3
Maximum	15
Mean	8.857
StdDev	3.823

Name: local Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

No.	Label	Count	Weight
1	ID	14	14.0
2	IM	14	14.0
3	PM	14	14.0
4	CHI	14	14.0
5	GAL	14	14.0

Name: class Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	0	11	11.0
2	1	59	59.0

Depois de feita a análise dos atributos e das estatísticas, iniciou-se o processo de classificação, utilizando a configurações instruídas no enunciado. Os resultados obtidos da árvore de decisão C4.5 foram:

- Número de folhas: 18
- Tamanho da árvore: 26

- Tempo de construção da árvore: 0.04s
- Porcentagem de classificação correta: 91.4286 %
- Kappa statistic: 0.6764
- Erro absoluto médio: 0.0929
- Raiz quadrada do erro absoluto médio: 0.2958
- Erro absoluto relativo: 34.0726 %
- Raiz quadrada do erro absoluto relativo: 81.1177 %
- Número total de instâncias: 70
- Matriz de confusão:

```
a  b   <-- classified as
8  3  |  a = 0
3 56 |  b = 1
```

Abaixo se exibe a árvore obtida:

```
J48 unpruned tree
------------------
temperatura_media <= 21.87
|   numero_dias_chuva <= 4
|   |   local = ID: 1 (3.0)
|   |   local = IM: 1 (3.0)
|   |   local = PM
|   |   |   numero_dias_chuva <= 3: 0 (1.0)
|   |   |   numero_dias_chuva > 3: 1 (2.0)
|   |   local = CHI
|   |   |   numero_dias_chuva <= 3: 1 (1.0)
|   |   |   numero_dias_chuva > 3: 0 (2.0)
|   |   local = GAL: 0 (3.0)
|   numero_dias_chuva > 4
|   |   local = ID: 1 (10.0)
|   |   local = IM
|   |   |   umidade_media <= 97.5: 1 (9.0)
|   |   |   umidade_media > 97.5: 0 (1.0)
|   |   local = PM: 1 (10.0)
|   |   local = CHI: 1 (10.0)
|   |   local = GAL: 1 (10.0)
temperatura_media > 21.87
|   local = ID: 1 (1.0)
|   local = IM: 0 (1.0)
|   local = PM: 0 (1.0)
|   local = CHI: 0 (1.0)
|   local = GAL: 0 (1.0)
```

Analisando a matriz de confusão observa-se que foi classificado corretamente 8 vezes que não havia a existência da espécie (a = 0) e 3 vezes de forma incorreta. Já da segunda linha da matriz observa-se que em 3 vezes foi classificado incorretamente a presença da espécie (b = 1) e em 56 vezes foi classificado corretamente.

Conforme solicitado no enunciado, apresenta-se a árvore de decisão na forma de regras SE ? ENTÃO classe é ? com 0 indicando classe a e 1 indicando classe b):

```
SE (temperatura_media <= 21.87 && numero_dias_chuva <= 4 && numero_dias_chuva <= 3 ) ENTÃO {
        ID=1 and IM=1 and GAL=0 and PM=0 and CHI=1
}
SE (temperatura_media <= 21.87 && numero_dias_chuva <= 4 && numero_dias_chuva > 3 ) ENTÃO {
        ID=1 and IM=1 and GAL=0 and PM=1 and CHI=0
}
SE (temperatura_media <= 21.87 && numero_dias_chuva > 4 && umidade_media <= 97.5 ) ENTÃO {
        ID=1 and IM=1 and GAL=1 and PM=1 and CHI=0
}
SE (temperatura_media <= 21.87 && numero_dias_chuva > 4 && umidade_media > 97.5 ) ENTÃO {
        ID=1 and IM=0 and GAL=1 and PM=1 and CHI=0
}
SE (temperatura_media > 21.87) ENTÃO {
        ID=1 and IM=0 and PM=0 and CHI=0 and GAL=0
}
```

Alterando os parâmetros minNunObj e unpruned (um de cada vez) obtém-se os resultados abaixo:

no caso de unpruned=false

- Número de folhas: 2
- Tamanho da árvore: 3
- Tempo de construção da árvore: 0.01s
- Porcentagem de classificação correta: 87.1429 %
- Kappa statistic: 0.4045
- Erro absoluto médio: 0.1679
- Raiz quadrada do erro absoluto médio: 0.3474
- Erro absoluto relativo: 61.6082 %
- Raiz quadrada do erro absoluto relativo: 95.2594 %
- Número total de instâncias: 70
- Matriz de confusão:

```
a  b   <-- classified as
4  7  |  a = 0
2 57  |  b = 1
```

Abaixo se exibe a árvore obtida:

```
J48 pruned tree
------------------
temperatura_media <= 21.87: 1 (65.0/7.0)
```

temperatura_media > 21.87: 0 (5.0/1.0)

no caso de minNumObj=2

- Número de folhas: 7
- Tamanho da árvore: 10
- Tempo de construção da árvore: 0.01s
- Porcentagem de classificação correta: 63% 90%
- Kappa statistic: 0.636
- Erro absoluto médio: 0.1154
- Raiz quadrada do erro absoluto médio: 0.2826
- Erro absoluto relativo: 42.3552%
- Raiz quadrada do erro absoluto relativo: 77.4978%
- Número total de instâncias: 70
- Matriz de confusão:

```
a  b   <-- classified as
8  3  |  a = 0
4 55 |  b = 1
```

Abaixo se exibe a árvore obtida:

```
J48 unpruned tree
------------------
temperatura_media <= 21.87
|   numero_dias_chuva <= 4
|   |   local = ID: 1 (3.0)
|   |   local = IM: 1 (3.0)
|   |   local = PM: 1 (3.0/1.0)
|   |   local = CHI: 0 (3.0/1.0)
|   |   local = GAL: 0 (3.0)
|   numero_dias_chuva > 4: 1 (50.0/1.0)
temperatura_media > 21.87: 0 (5.0/1.0)
```

Analisando os resultados em comparação com o obtido na configuração anterior, quando se aumenta o MinNumObj para dois, a árvore se torna mais simples havendo uma perda na performance. Visto que essa perda é pequena, essa configuração dos parâmetros se torna interessante de ser aplicada. Ao colocar unpruning como false, o desempenho cai muito em comparação com a primeira configuração, devendo se utilizar esse parâmetro somente quando se necessita algo extremamente simples.

Da análise do parágrafo anterior, deduz-se que quando a árvore é “unpruned=true” ela não passa por um processo de “poda” e por isso é maior. Ao colocar essa opção como false, após a construção da árvore executa-se o passo da poda que verifica a existência de partes a serem removidas sem causar uma grande perda da performance diminuindo o risco de overfitting. O parâmetro MinNumObj configura o valor mínimo de instâncias em uma folha.

Obtendo outro conjunto de dados com atributos categóricos disponível em https://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/contact-lenses.arff se construiu a árvore rodando a classificação para diversas configurações dos parâmetros, obtendo os melhores resultados com **unpruned=false** e **minNumObj=3** conseguindo os resultados abaixo:

- Número de folhas: 4
- Tamanho da árvore: 7
- Tempo de construção da árvore: 0.01s
- Porcentagem de classificação correta: 21% 87.5%
- Kappa statistic: 0.7895
- Erro absoluto médio: 0.1444
- Raiz quadrada do erro absoluto médio: 0.2944
- Erro absoluto relativo: 38.2353%
- Raiz quadrada do erro absoluto relativo: 67.4061%
- Número total de instâncias: 24
- Matriz de confusão:

```
a  b  c   <-- classified as
5  0  0   |  a = soft
0  4  0   |  b = hard
1  2 12   |  c = none
```

Abaixo se exibe a árvore obtida:

```
J48 pruned tree
------------------
tear-prod-rate = reduced: none (12.0)
tear-prod-rate = normal
|   astigmatism = no: soft (6.0/1.0)
|   astigmatism = yes
|   |   spectacle-prescrip = myope: hard (3.0)
|   |   spectacle-prescrip = hypermetrope: none (3.0/1.0)
```