

**Escola Politécnica de Pernambuco**

*Especialização em Ciência de Dados e Analytics*

**Trabalho da disciplina: Introdução à Ciência de Dados**

Grupo:

* Adriana Rodrigues da Silva
* Diego Lages dos Santos
* Laís Ingrid Soares de Souza Vidoto
* Marcus Vinícius Raposo Azevedo

Prof. Dr. Alexandre Maciel

Sumário

[Dinâmica 01 - Com os grupos montados na aula passada, vamos: 4](#_Toc4540767)

[Dinâmica 02 - Utilize a base de dados de mamografia para realizar a análise descritiva de dados 5](#_Toc4540768)

[Dinâmica 03 - Utilize as bases de dados do Weka para analisar uma técnica de classificação ou de agrupamento e descreva suas considerações. 8](#_Toc4540769)

[Técnica de Classificação 8](#_Toc4540770)

[Resultados 9](#_Toc4540771)

[Técnica de Agrupamento 10](#_Toc4540772)

[Resultados 12](#_Toc4540773)

[Anexos 14](#_Toc4540774)

[A - Configuração do RandomForest 14](#_Toc4540775)

[B – Configuração NaiveBayes 14](#_Toc4540776)

[C – Configuração MultilayerPerceptron 15](#_Toc4540777)

[D – Configuração EM 16](#_Toc4540778)

[E – Configuração SimplesKMeans 16](#_Toc4540779)

Índice de figuras

[Figura 1 - Registros inconsistêntes de BI-RADS 5](#_Toc4540780)

[Figura 2 - Registros inconsistêntes (AGE) 5](#_Toc4540781)

[Figura 3 - Quadro com a distribuição das frequências 6](#_Toc4540782)

[Figura 4 - Histograma do AGE 6](#_Toc4540783)

[Figura 5 - Gráfico de dispersão (Idade x Severidade) 6](#_Toc4540784)

[Figura 6 - Tabela de Médias 7](#_Toc4540785)

[Figura 7 - Gráfico de dispersão 7](#_Toc4540786)

[Figura 8 - Boxplot (AGE) 7](#_Toc4540787)

[Figura 9 - Resultado RandomForest 9](#_Toc4540788)

[Figura 10 - Resultado NaiveBayes 9](#_Toc4540789)

[Figura 11 - Resultado MultilayerPerceptron 9](#_Toc4540790)

[Figura 12 - Resultado algoritmo EM 12](#_Toc4540791)

[Figura 13 - Resultado algoritmo SimplesKMeans 12](#_Toc4540792)

[Figura 14 - Resultado algoritmo SimplesKMeans, com 4 grupos 12](#_Toc4540793)

[Figura 15 - Visualização dos agrupamentos por atributos 12](#_Toc4540794)

[Figura 16 - Configuração do algoritmo RandomForest 14](#_Toc4540795)

[Figura 17 - Configuração do algoritmo NaiveBayes 14](#_Toc4540796)

[Figura 18 - Configuração do algoritmo MultilayerPerceptron 15](#_Toc4540797)

[Figura 19 - Configuração do algoritmo EM 16](#_Toc4540798)

[Figura 20 - Configuração do algoritmo SimplesKMeans 16](#_Toc4540799)

# Dinâmica 01 - Com os grupos montados na aula passada, vamos:

1. Discutir problemas

2. Identificar stakeholders

3. Indicar possíveis bases de dados

4. Apresentar descobertas anteriores

**Problema**: A empresa XYZ não sabe ao certo quantos clientes possui. Existem divergências acentuadas no número disponibilizado pelo Comercial, pelo Financeiro e por Operação. Ao que tudo indica essa diferença não é resultante de uma simples divergência conceitual quanto a definição de cliente, mas sim de erros humanos e sistêmicos no registro da informação em cada unidade, e a falta de controle no uso de licenças para alguns sistemas comercializados pela empresa. Essa questão, aparentemente simples, gera impactos negativos em toda organização uma vez que lança dúvida quanto ao cumprimento da visão da empresa que é atingir a carteira de 5.000 clientes recorrentes até o final de 2020, na medida que os números levantados giram em torno de 1.200 a 1.400, o que gera dúvidas quanto a forma de coleta e manutenção desse número.

**Stakeholders**: Entre os stakeholders envolvidos podemos citar por área:

* Diretoria: CEO, Diretor Adm-financeiro, Diretor comercial, Gerente de operações;
* Financeiro: Supervisor financeiro, analista financeira responsável pelo registro dos contratos na ferramenta de gestão financeira da empresa;
* Comercial: Gerente comercial, Assistente administrativo responsável pelo registro dos novos clientes e clientes cancelados.

**Possíveis bases de dados**: CRM Dynamics (Comercial), Sistema Pirâmide (Gestão financeira), Sistema comercial (Sistema utilizado por Operações para controlar a licença de alguns softwares comercializados pela empresa, Planilha excel utilizada pelo comercial para controlar os cancelamentos de clientes e a aquisição de novos clientes.

**Descobertas anteriores**:

# Dinâmica 02 - Utilize a base de dados de mamografia para realizar a análise descritiva de dados

1. Distribuição de frequência

• Limite inferior, superior, ponto médio, frequência absoluta e relativa

2. Visualização de Dados

• Histograma, gráfico de dispersão

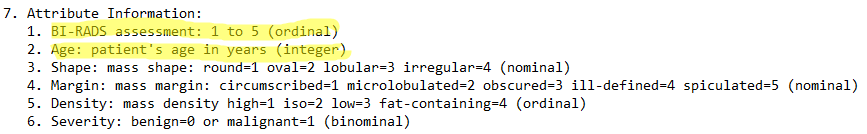
3. Medidas de Resumo

• Médias, diagrama de dispersão, box plot

**Distribuição da frequência (age por BI-RADS assessment).**

• Limite inferior, superior, ponto médio, frequência absoluta e relativa

A primeira atividade foi consultar o dicionário de dados (arquivo: [mammographic\_masses.names](http://archive.ics.uci.edu/ml/machine-learning-databases/mammographic-masses/mammographic_masses.names)) para entendermos mais sobre cada coluna do banco de dados. Logo em seguida verificamos que existiam registros na base que não estavam em conformidade com as restrições definidas para os campos: BI-RADS e Age, conforme tabela abaixo:



Nesse sentido, iniciamos a limpeza da base de dados nos campos BI-RADS e Age, excluindo os registros inconsistentes, conforme as tabelas abaixo:

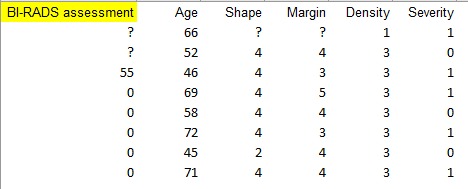


Figura 1 - Registros inconsistêntes de BI-RADS

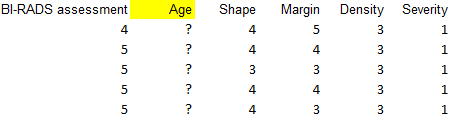


Figura 2 - Registros inconsistêntes (AGE)

Após realizarmos a higienização dos campos BI-RADS e AGE, utilizamos o Ms-Excel e criamos uma tabela dinâmica para calculamos as frequências do campo AGE agrupando pelo campo BI-RADS, conforme resultado abaixo:

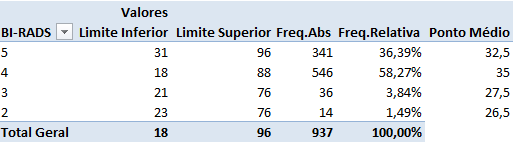


Figura 3 - Quadro com a distribuição das frequências

Para mais informações consultar o arquivo anexo: Questão-02.xlsx.

**Visualização de Dados**

• Histograma, gráfico de dispersão

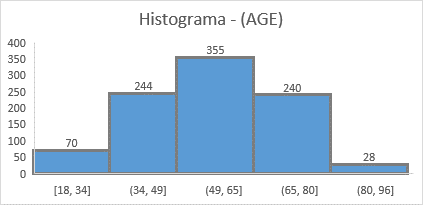


Figura 4 - Histograma do AGE

Figura 5 - Gráfico de dispersão (Idade x Severidade)

Para mais informações consultar o arquivo anexo: Questão-02.xlsx.

**Medidas de Resumo**

• Médias, diagrama de dispersão, box plot

|  |  |
| --- | --- |
| **IDADE** | |
| Média | 55,39 |
| Mediana | 57,00 |
| Ponto médio | 39,00 |
| Moda | 59,00 |

Figura 6 - Tabela de Médias

Figura 7 - Gráfico de dispersão

Para calculo do LI (limite inferior) e LS (limite superior) utilizamos as seguintes fórmulas:

LI (limite inferior): 1oQuartil -1,5\*(3oQuartil - 1oQuartil)

LS (limite superior): 3oQuartil +1,5\*(3oQuartil - 1oQuartil)

Figura 8 - Boxplot (AGE)

Para mais informações consultar o arquivo anexo: Questão-02.xlsx.

# Dinâmica 03 - Utilize as bases de dados do Weka para analisar uma técnica de classificação ou de agrupamento e descreva suas considerações.

## Técnica de Classificação

Para a técnica de classificação escolhemos de forma aleatória a base diabetes.arff. Esta base foi cedida em maio de 1990 pela “*National Institute of Diabetes and Digestive and Kidney Diseases*”, e a variável de valor binário investigada é se o paciente apresenta sinais de diabetes de acordo com critérios definidos pela Organização Mundial de Saúde (0 para teste negativo e 1 para valor positivo).

Várias restrições foram utilizadas na criação desse banco de dados. Em particular, todos os pacientes aqui são do sexo feminino com menos 21 anos de idade, da herança indígena Pima e vivem perto de Phoenix, Arizona, EUA.

Abaixo seguem algumas informações sobre a base de dados.

**Número de instâncias**: 768

**Número de atributos**: 7 atributos e 1 classe (target)

**Atributos**: (todos numéricos)

1. Number of times pregnant

2. Plasma glucose concentration a 2 hours in an oral glucose tolerance tes

3. Diastolic blood pressure (mm Hg)

4. Triceps skin fold thickness (mm)

5. 2-Hour serum insulin (mu U/ml)

6. Body mass index (weight in kg/(height in m)^2)

7. Diabetes pedigree function

8. Age (years)

9. Class variable (0 or 1)

A primeira tarefa foi verificar se existiam valores faltantes nos atributos para aplicarmos alguma técnica de limpeza nos dados. Constatamos que não existiam valores faltantes na base, nesse sentido resolvemos aplicar alguns algoritmos de classificação com sua configuração padrão. Segue abaixo breve relato sobre os parâmetros utilizados em cada um dos algoritmos.

* **Técnica árvore de decisão com o algoritmo RandomForest**: Utilizamos os parâmetros padrões apresentados pela ferramenta, batchsize = 100, numIterations = 100, além de aplicarmos um cross-validation folds de 10. (Conforme Anexo A).
* **Técnica Bayes com o algoritmo NaiveBayes**: Utilizamos os parâmetros default apresentados pela ferramenta, batchsize = 100, além de aplicarmos um cross-validation folds de 10. (Conforme Anexo B).
* **Técnica de redes neurais com o algoritmo Multi-layer perceptron**: Utilizamos os parâmetros default apresentados pela ferramenta, batchsize = 100, training time = 500, além de um cross-validation folds de 10 (Conforme Anexo C)

### Resultados

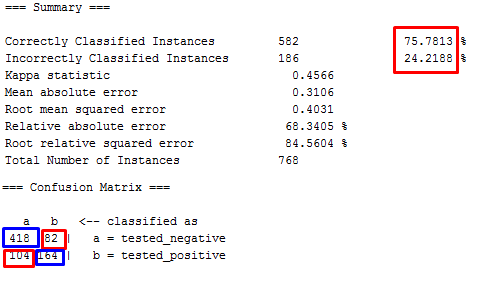


Figura - Resultado RandomForest

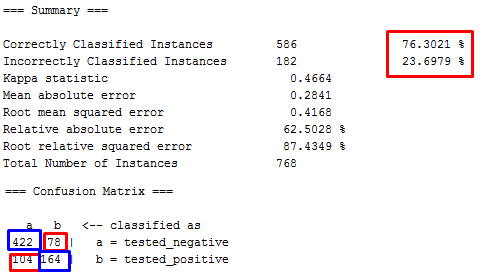


Figura - Resultado NaiveBayes

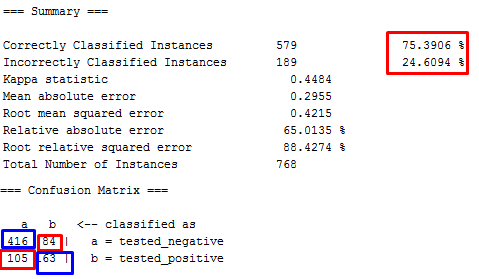


Figura - Resultado MultilayerPerceptron

Para avaliar o resultado dos algoritmos, resolvemos utilizar não somente a acurácia, mas sobretudo realizar uma análise na matriz de confusão, uma vez que para esse tipo de problema (identificação ou não de uma doença: diabetes) resultados falso positivo ou falso negativos tem um grande impacto negativo na vida dos pacientes que recebem a notícia.

Ao analisarmos o resultado, o que nos chamou a atenção não foi o fato do algoritmo NaiveBayes ter alcançado a melhor acurácia (76,3021%) em relação ao demais (RandomForest = 75,7813%; Multilayer Perceptron = 75,3906%), mas sim, de também ter apresentado a melhor taxa de verdadeiro positivos, verdadeiro negativos além de obter uma menor taxa de erros (falso positivos e falsos negativos), como podemos observar no resumo abaixo.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algoritmos | ACERTOS | | ERROS | |
| Verdadeiro  Positivo | Verdadeiro Negativo | Falso  Positivo | Falso  Negativo |
| NaiveBayes | 164 | 422 | 78 | 104 |
| RandomForest | 164 | 418 | 82 | 104 |
| Multilayer Perceptron | 163 | 416 | 84 | 105 |

Obs.: A título de curiosidade, resolvemos aplicar o método PCA em cada algoritmo e constatamos que o desempenho foi inferior a execução dos mesmos algoritmos sem executar o método PCA. Atribuímos esse fato a constatação de que o método PCA é normalmente utilizado para reduzir dimensionalidade, quando a quantidade de atributos é muito elevada, através da seleção dos atributos que mais contribuem para a predição do target. Nossa base de dados possuía apenas 8 atributos, e o método PCA “removeu” apenas 1 desses atributos, dessa maneira pudemos constatar na prática a falta de efetividade deste método em base de dados simples com poucos atributos.

Após aplicarmos os algoritmos em sua configuração padrão, escolhemos o algoritmo que apresentou melhor desempenho, NaiveBayes, e realizamos alguns ajustes em seus parâmetros visando a otimização do modelo, eis o resultado:

* Não houve melhora significativa ao aumentar ou diminuir o cross-validation. O melhor resultado foi obtido com o valor padrão 10. (Testamos: 7,8,9,11,12,13,14,15).
* Não houve melhora significativa ao aumentar o número de batchsize para 500 ou 1.000. O melhor resultado foi obtido com o valor padrão 100.
* Houve piora significativa ao usar um percentual para dividir o conjunto de treino e teste (70%) em substituição do método de cross-validation = 10.

Como conclusão, acreditamos que os valores padrões dos algoritmos já são estabelecidos com o objetivo de maximizar o resultado, o que não invalida de maneira nenhuma a tentativa de ajuste e otimização dos parâmetros. Adicionalmente, concluímos que é de suma importância estudar os demais algoritmos de classificação que não usamos nesse trabalho, procurando identificar as vantagens, desvantagens e situações de uso de cada um.

## Técnica de Agrupamento

Para a técnica de agrupamento, também escolhemos uma base de dados aleatória, no caso a labor.arff. Essa base foi cedida em novembro de 1988 por “Stan Matwin, Computer Science Dept, University of Ottawa”. Os dados incluem todos os acordos coletivos alcançados no setor de negócios e serviços pessoais para os moradores com pelo menos 500 membros (professores, enfermeiros, funcionários da universidade, polícia, etc) no Canadá em 87 e no primeiro trimestre de 88.

Abaixo seguem algumas informações sobre a base de dados.

**Número de instâncias**: 57

**Número de atributos**: 16 + 1 classe (good or bad)

**Atributos**:

1. dur: duration of agreement [1..7]
2. wage1.wage : wage increase in first year of contract [2.0 .. 7.0]
3. wage2.wage : wage increase in second year of contract [2.0 .. 7.0]
4. wage3.wage : wage increase in third year of contract [2.0 .. 7.0]
5. cola : cost of living allowance [none, tcf, tc]
6. hours.hrs : number of working hours during week [35 .. 40]
7. pension : employer contributions to pension plan [none, ret\_allw, empl\_contr]
8. stby\_pay : standby pay [2 .. 25]
9. shift\_diff : shift differencial : supplement for work on II and III shift [1 .. 25]
10. educ\_allw.boolean : education allowance [true false]
11. holidays : number of statutory holidays [9 .. 15]
12. vacation : number of paid vacation days [ba, avg, gnr]
13. lngtrm\_disabil.boolean : employer's help during employee longterm disability [true , false]
14. dntl\_ins : employers contribution towards the dental plan [none, half, full]
15. bereavement.boolean : employer's financial contribution towards the covering the costs of bereavement [true , false]
16. empl\_hplan : employer's contribution towards the health plan [none, half, full]

A primeira tarefa foi verificar se existiam valores faltantes nos atributos para aplicarmos alguma técnica de limpeza nos dados. Constatamos que existiam vários atributos com valores faltantes na base, nesse sentido utilizamos a opção de pré-processamento: “*ReplaceMissingValues*”, que substitui todos os valores ausentes de atributos nominais e numéricos em um conjunto de dados com os modos e médias dos dados de treinamento. Após “ajustarmos” os valores faltantes, sob orientação do professor Alexandre, removemos o atributo 17 (relacionado a classe) antes de aplicar dois algoritmos de agrupamento com sua configuração padrão. Segue abaixo breve relato sobre os parâmetros utilizados em cada um dos algoritmos.

* **Técnica de clusterização através do algoritmo EM (Expectation maximisation):** Utilizamos os parâmetros padrões apresentados pela ferramenta, maxiterations = 100, numFolds = numKmeansRuns = 10. Além de usar um conjunto para treinamento. (Conforme Anexo D)
* **Técnica de clusterização através do algoritmo SimplesKmeans:** Utilizamos os parâmetros padrões apresentados pela ferramenta, maxiterations = 500, numClusters = 2. Além de usar um conjunto para treinamento. (Conforme Anexo E).

### Resultados

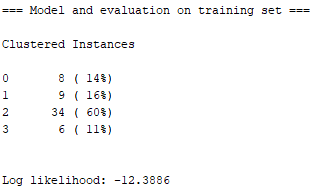


Figura - Resultado algoritmo EM

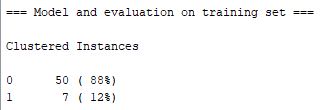


Figura - Resultado algoritmo SimplesKMeans

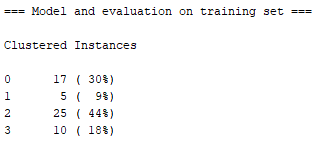


Figura - Resultado algoritmo SimplesKMeans, com 4 grupos

Ao executarmos o algoritmo SimplesKMeans com 2 clusters definidos como parâmetro, podemos observar que a distribuição se concentra em um dos grupos (grupo 0 com 80%). De maneira similar, ao executarmos o algoritmo do EM (Expectation maximisation), o número de agrupamentos otimizado foi automaticamente definido para 4, com um grupo se destacando pela grande quantidade de componentes (grupo 2 com 60%). Ao executarmos o algoritmo do SimplesKMeans com 4 clusters definidos como parâmetros, obtemos uma distribuição um pouco mais equilibrada em relação ao algoritmo EM.

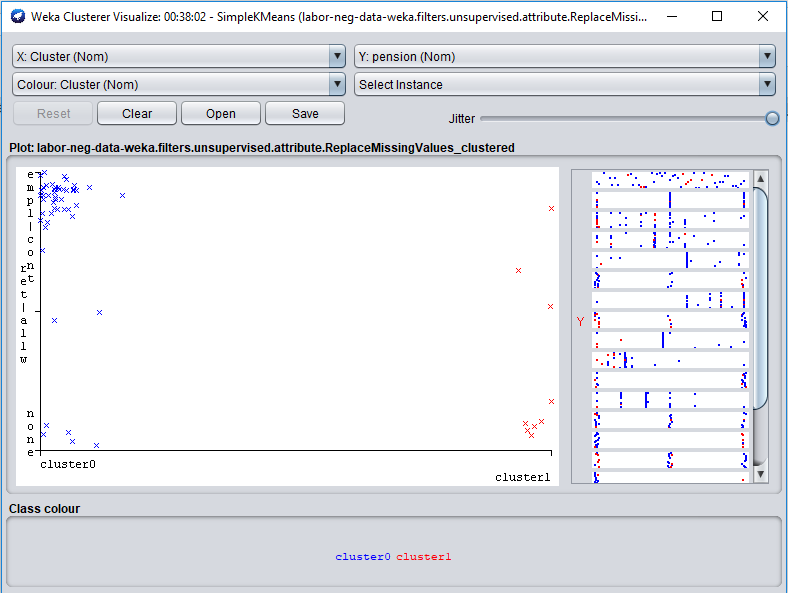


Figura - Visualização dos agrupamentos por atributos

Como conclusão, podemos destacar que o estudo aprofundado das características da base, nos problemas relacionados a clusterização (agrupamento) são de fundamental importância, uma vez que nesse tipo de problema não existe um TARGET a ser atingido, como nos problemas de classificação, mas sim, o estudo das características comuns de cada agrupamento a fim de obtermos insights para o negócio.

Infelizmente não conseguimos realizar um estudo detalhado sobre o significado e possíveis relações dos grupos sinalizados pelos algoritmos de clusters utilizados. Dessa forma não foi possível determinar qual algoritmo de agrupamento possibilitou que mais insights fossem gerados.

Obs.: Adicionalmente, aplicamos o método PCA com o objetivo de reduzir a dimensionalidade da base de dados através da redução do número de atributos, pois imaginávamos que dessa forma seria mais fácil realizar o estudo das correlações entre os atributos, contudo, para nossa surpresa, o PCA não removeu nenhuma coluna.

# Anexos

## A - Configuração do RandomForest

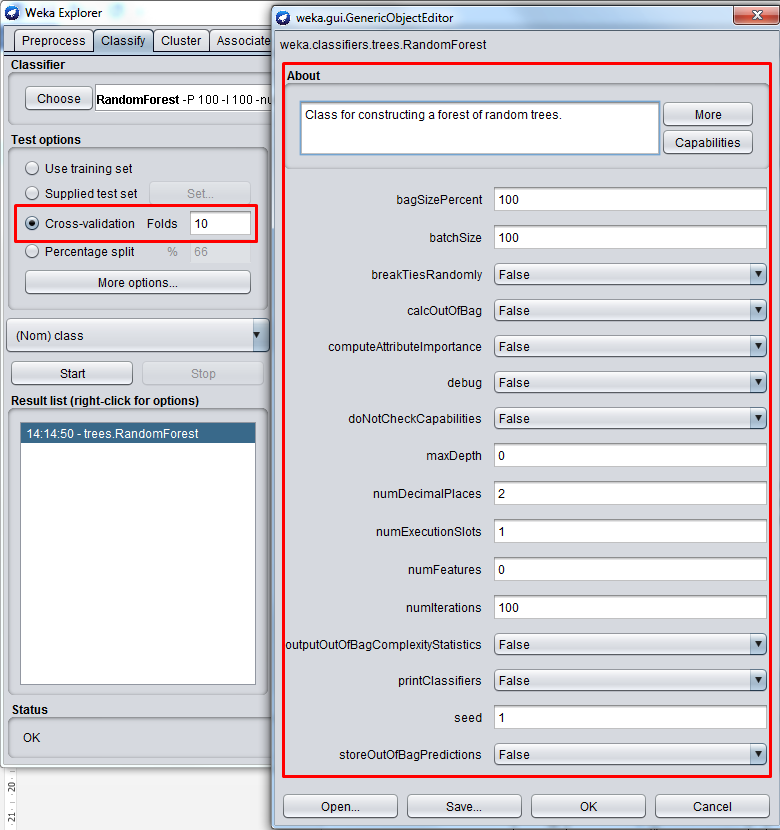


Figura - Configuração do algoritmo RandomForest

## B – Configuração NaiveBayes

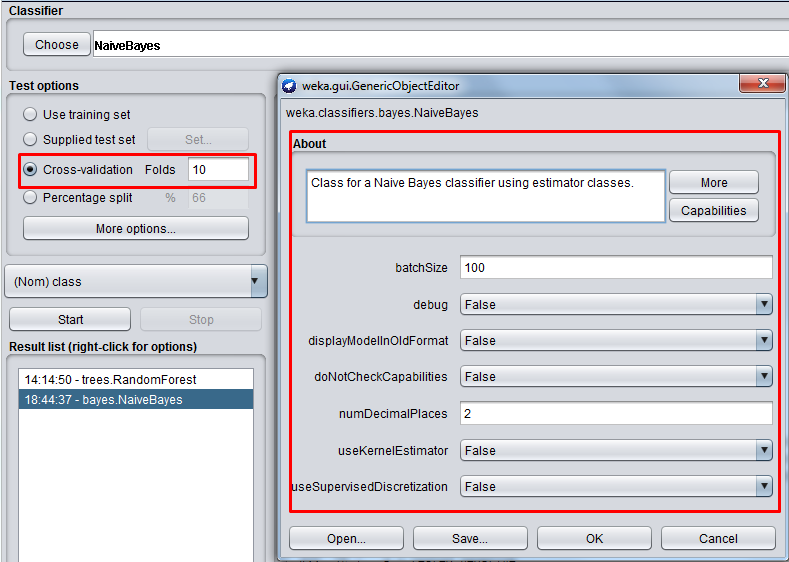


Figura - Configuração do algoritmo NaiveBayes

## C – Configuração MultilayerPerceptron

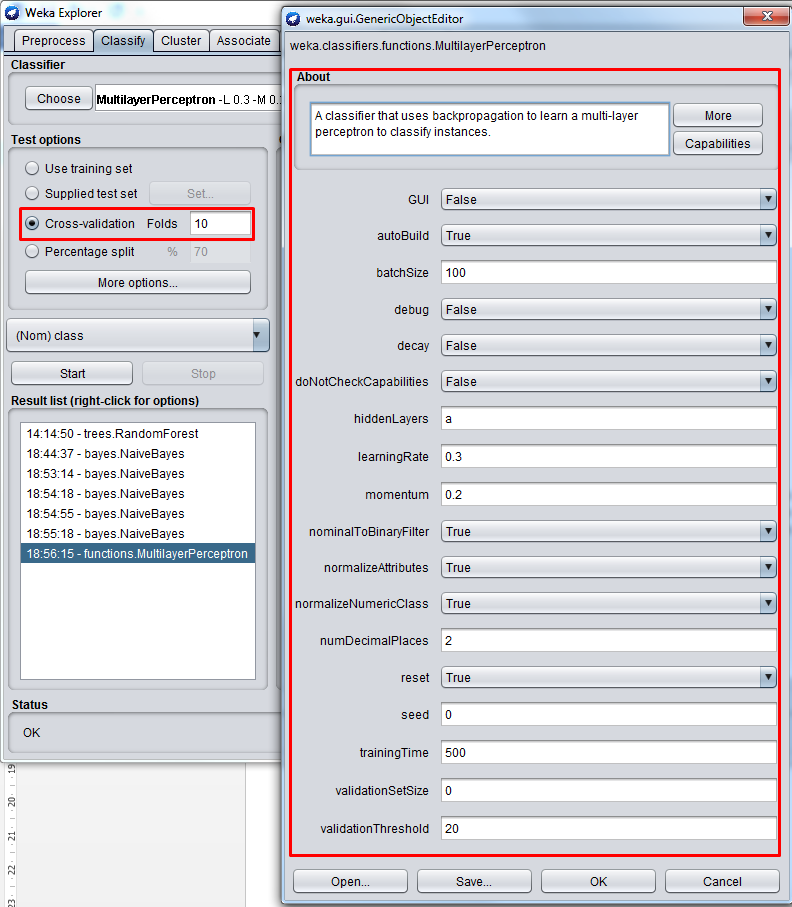


Figura - Configuração do algoritmo MultilayerPerceptron

## D – Configuração EM

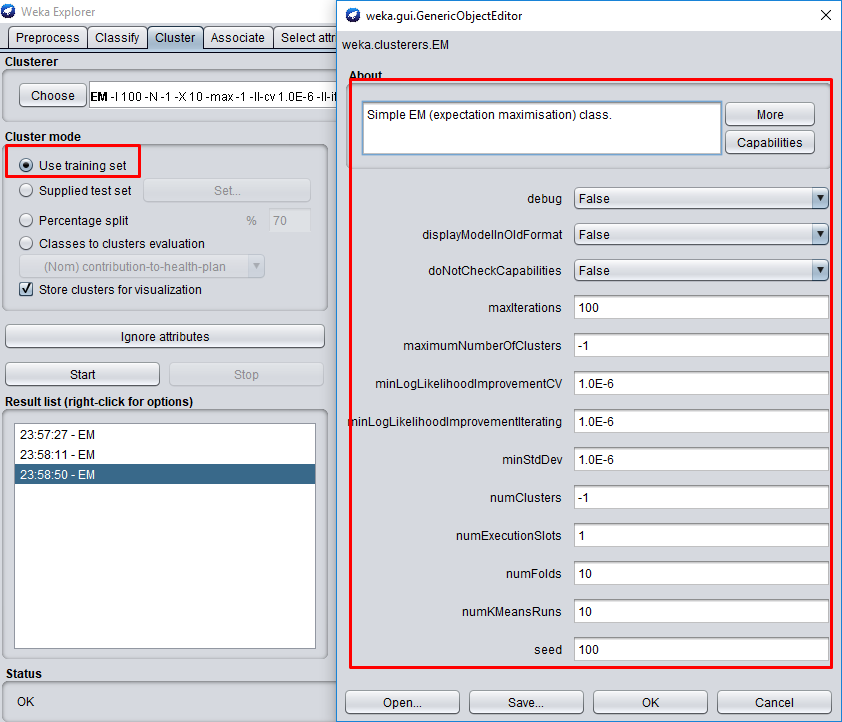


Figura - Configuração do algoritmo EM

## E – Configuração SimplesKMeans

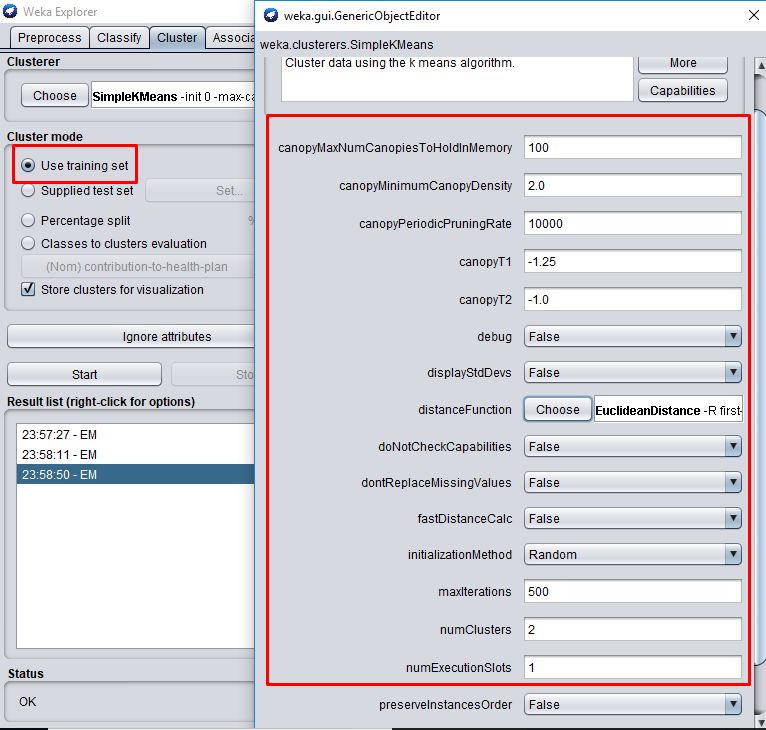


Figura - Configuração do algoritmo SimplesKMeans