Departamento de Engenharia Informática e Sistemas

Licenciatura em Engenharia Informática

Sistemas de Informação II

Isabel Castro 2018013160 & Maria Marcos 2018018386

DATA MINING

2020/2021

Índice

[Introdução 2](#_Toc62416223)

[Classificadores Decision Tree, Naive Bayes e Neural Network 3](#_Toc62416224)

[Alvo: continente 3](#_Toc62416225)

[Alvo: quadrante geográfico 7](#_Toc62416226)

[Alvo: língua 9](#_Toc62416227)

[Alvo: religião 10](#_Toc62416228)

[Técnicas de clustering 12](#_Toc62416229)

[Regras de associação 20](#_Toc62416230)

[20](#_Toc62416231)

[Regras de associação 21](#_Toc62416232)

[22](#_Toc62416233)

[Regras de associação 23](#_Toc62416234)

[Regras de associação 24](#_Toc62416236)

[25](#_Toc62416237)

[Conclusão 26](#_Toc62416238)

## Introdução

No âmbito da unidade curricular de Sistemas de Informação II foi-nos solicitada a realização de um trabalho em Data-Mining.

Esta ferramenta consiste na mineração de dados, isto é uma funcionalidade que agrega e organiza dados, encontrando neles padrões, associações, mudanças e anomalias relevantes. Este trabalho teve como base o ficheiro flags.csv que contém dados e descrições relativas às bandeiras dos países do mundo.

Assim, no inicio do trabalho fizemos o *import* e *backup* deste ficheiro no Microsoft SQL Server Management Studio em forma de base de dados, para depois ser mais fácil importá-lo para o Microsoft Visual Studio.

Perante esta base de dados, foi-nos solicitada a implementação de 4 tarefas de classificação, tendo como targets (alvos) o elementos : continente, quadrante geográfico, língua e religião.

O objetivo foi perceber se através das cores das bandeiras de cada país este conseguia descobrir ou ver associações corretas entre estes e os diferentes alvos.

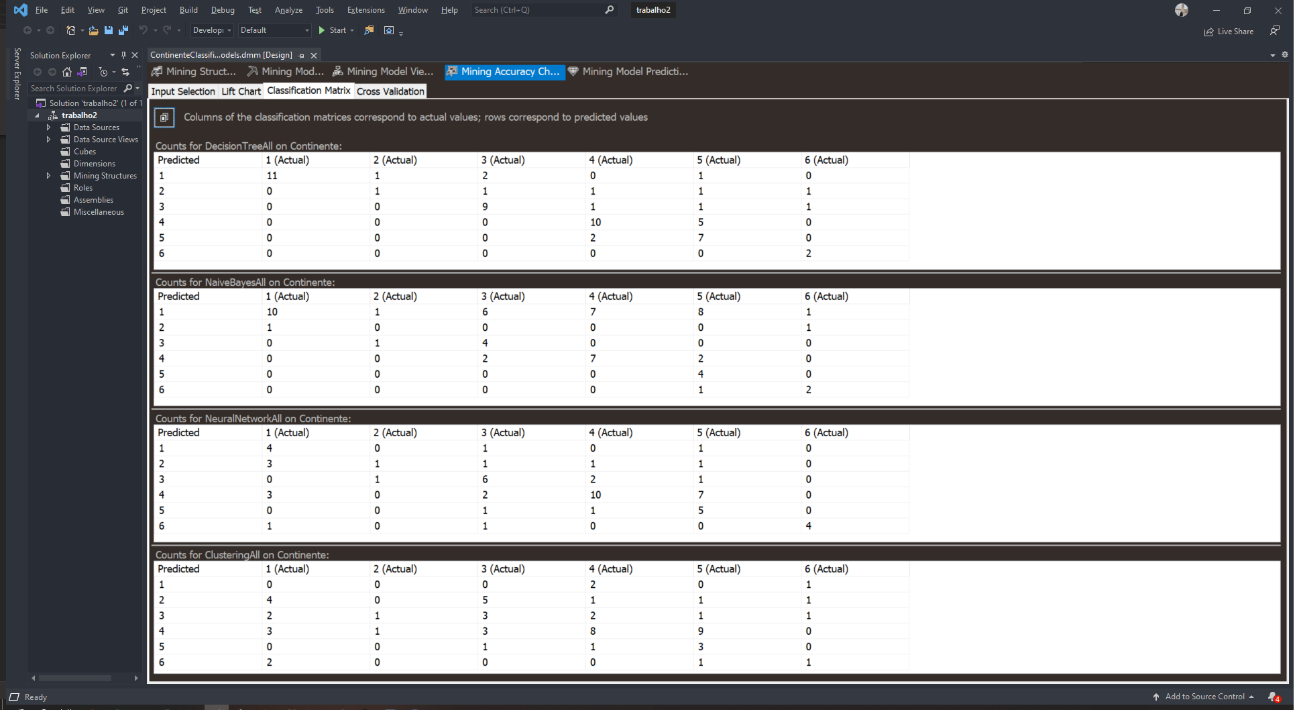
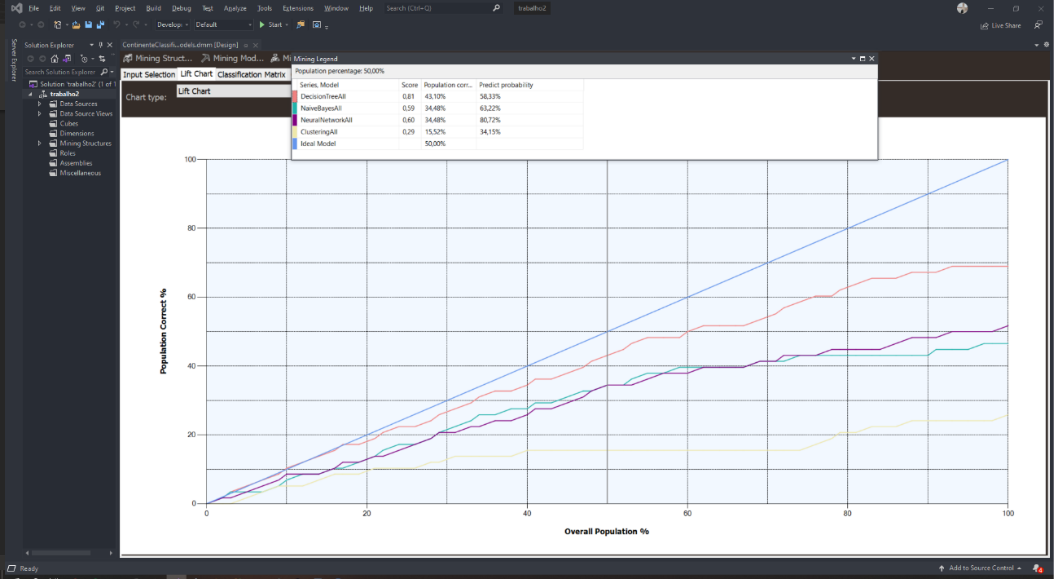
De modo a obtermos estes resultados foram utilizadas técnicas de Decision Tree, Clustering, NeuralNetwork, NativeBayes, Association Rules e ainda Data mining Lift Chart for Mining Structure.

Para conseguirmos ter mais termos de comparação realizamos diferentes testes:

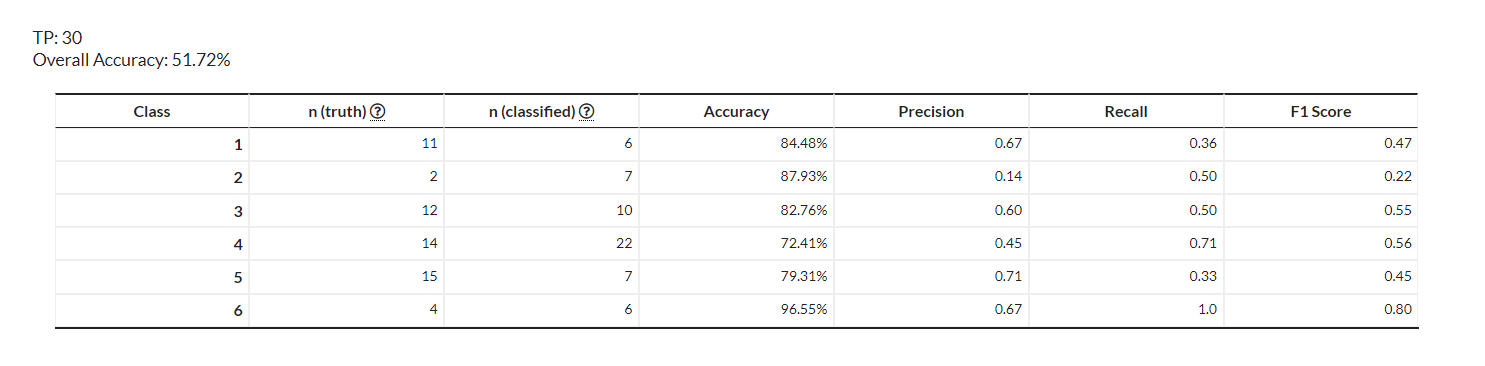
* Todas as entradas
* Apenas as bandeiras
* Bandeiras + atributo x ou y

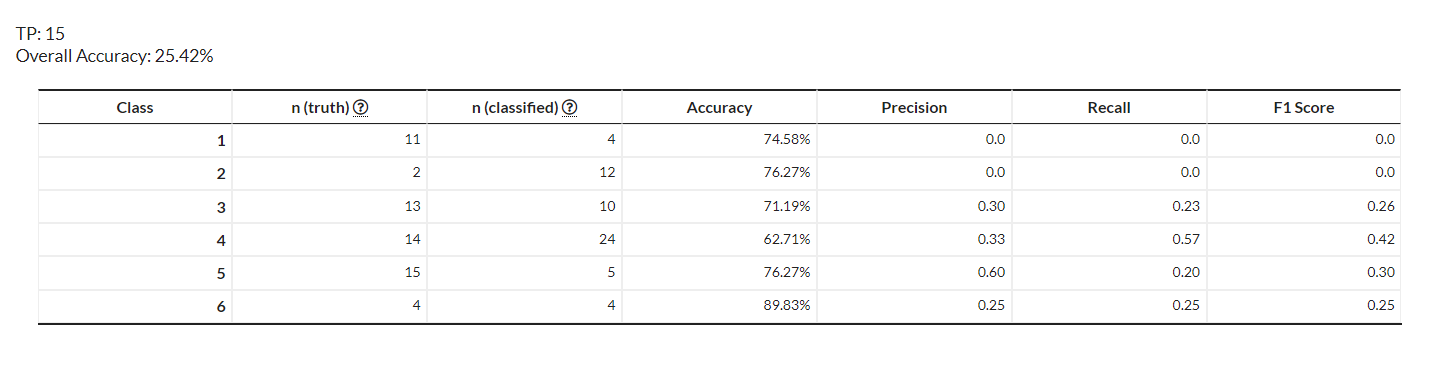
## Classificadores Decision Tree, Naive Bayes e Neural Network

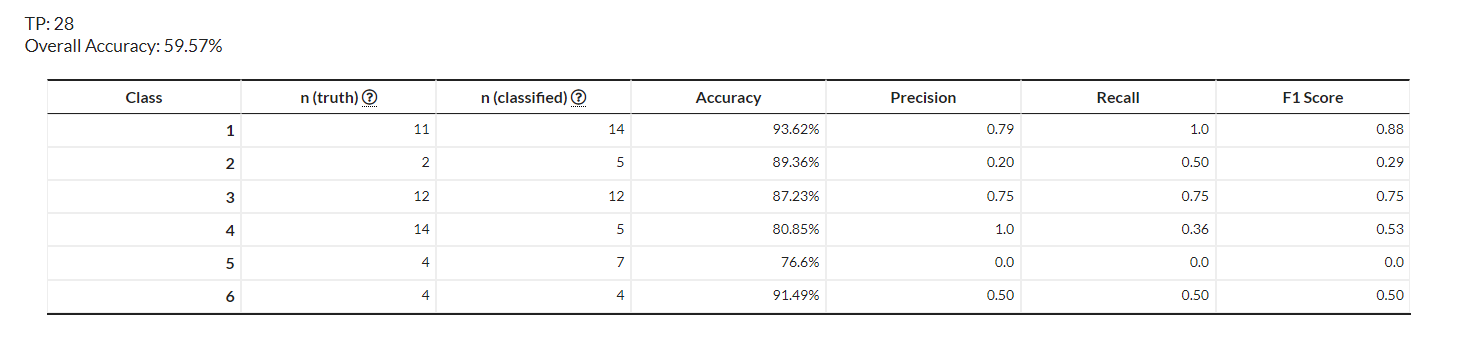
### Alvo: continente

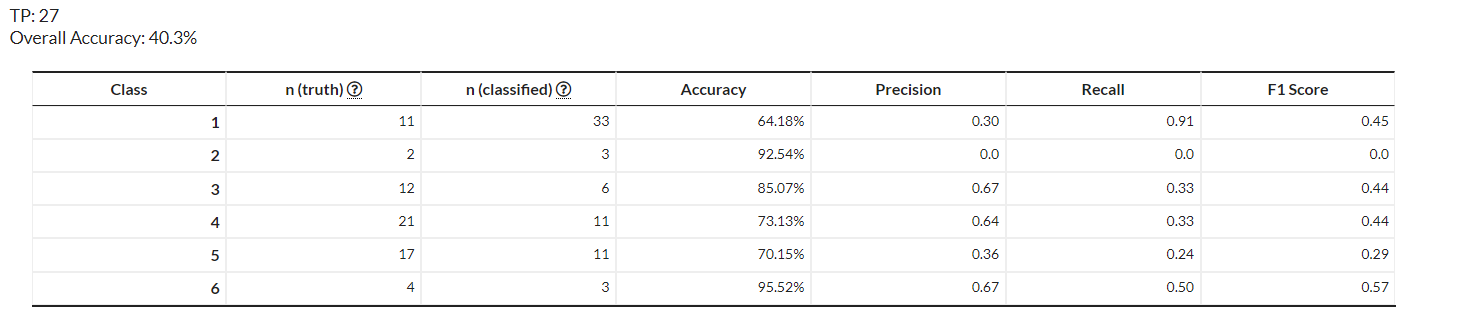
* Predicit: continente
* Key: Pais
* Inputs: Características das bandeiras + religião +quadrante + população

#### Cálculo das métricas

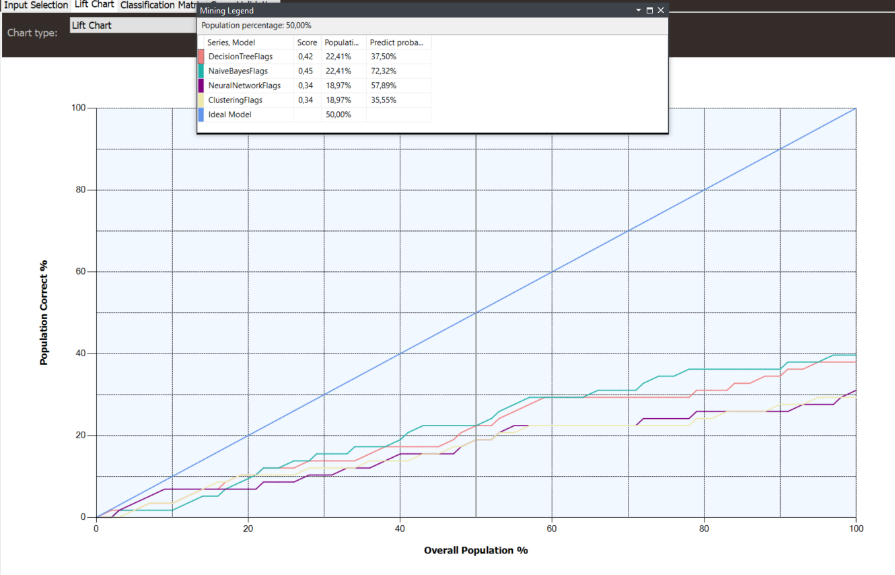


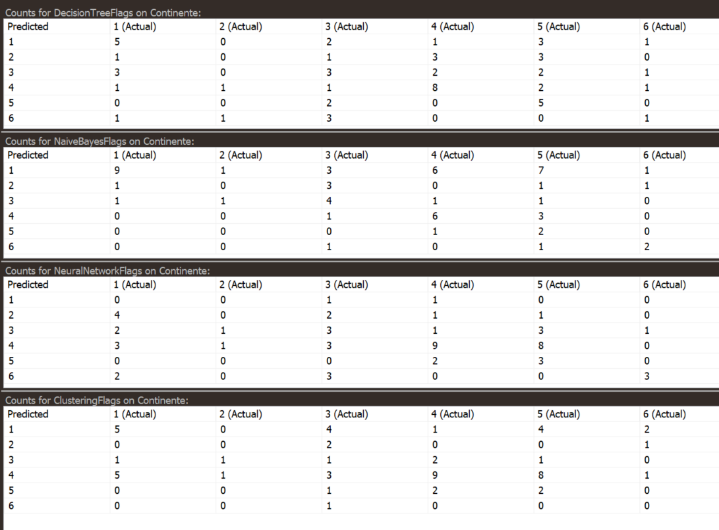




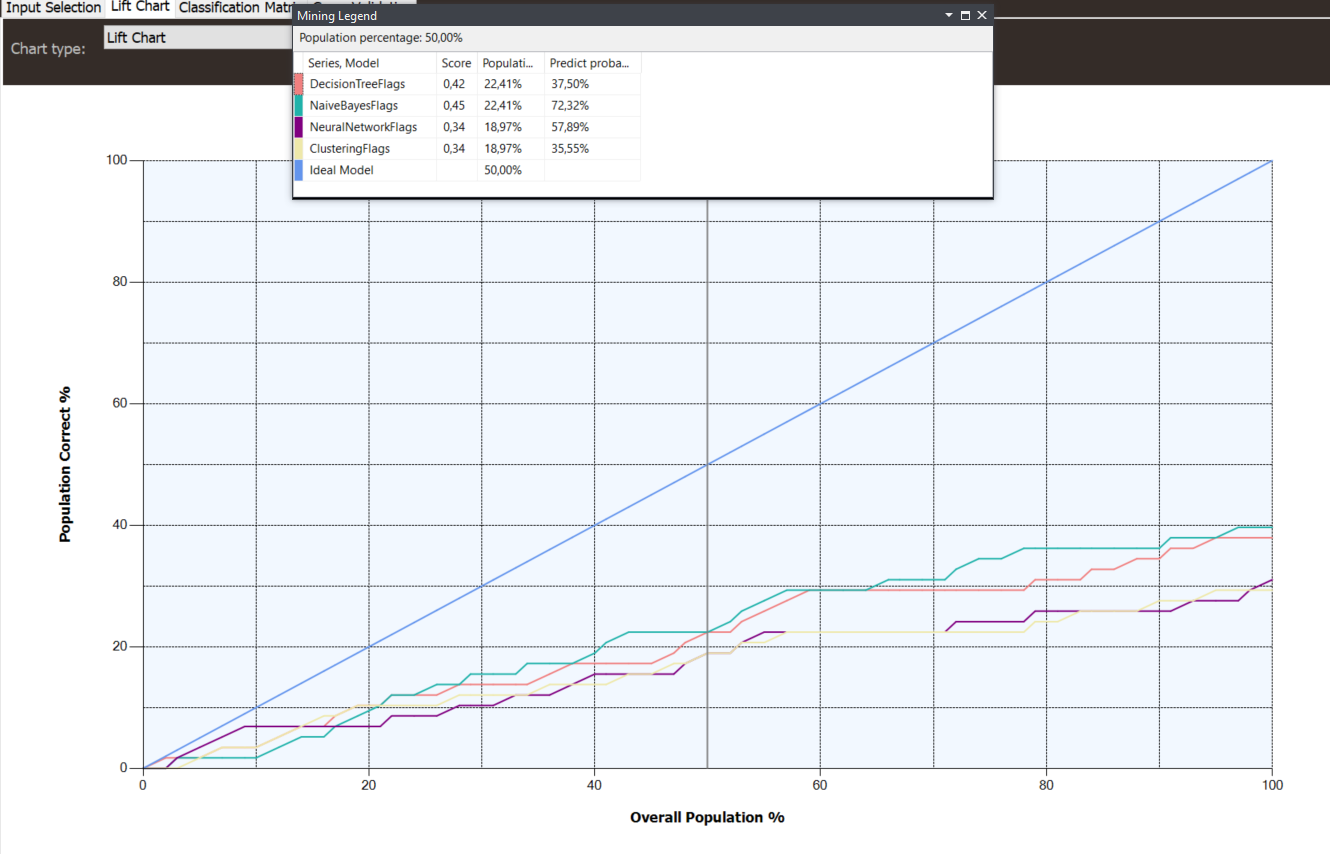
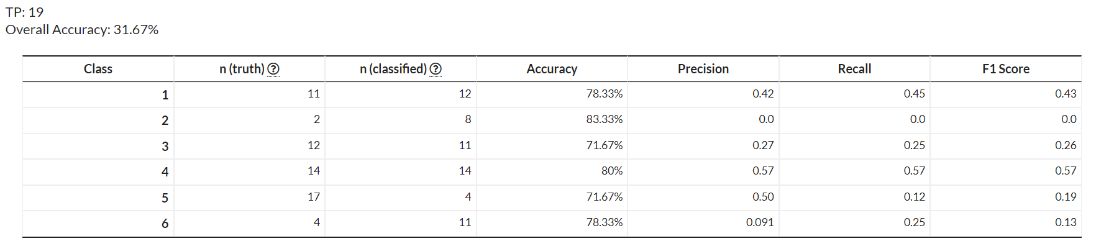


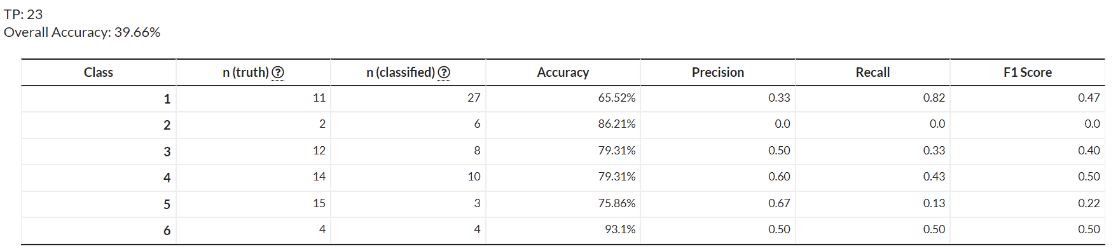
#### Conclusões

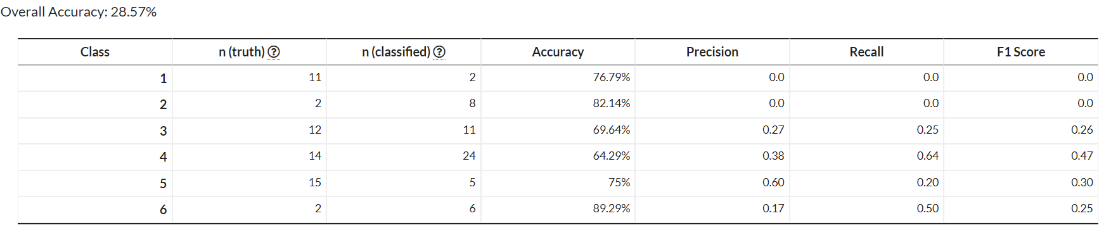
* Predicit: continente
* Key: Pais
* Inputs: Características das bandeiras + população

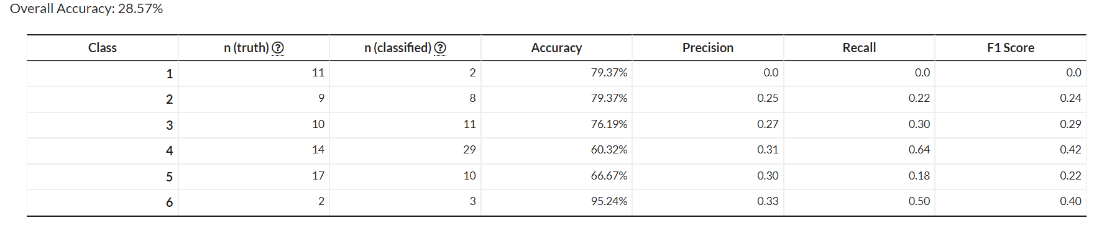


#### Cálculo das métricas

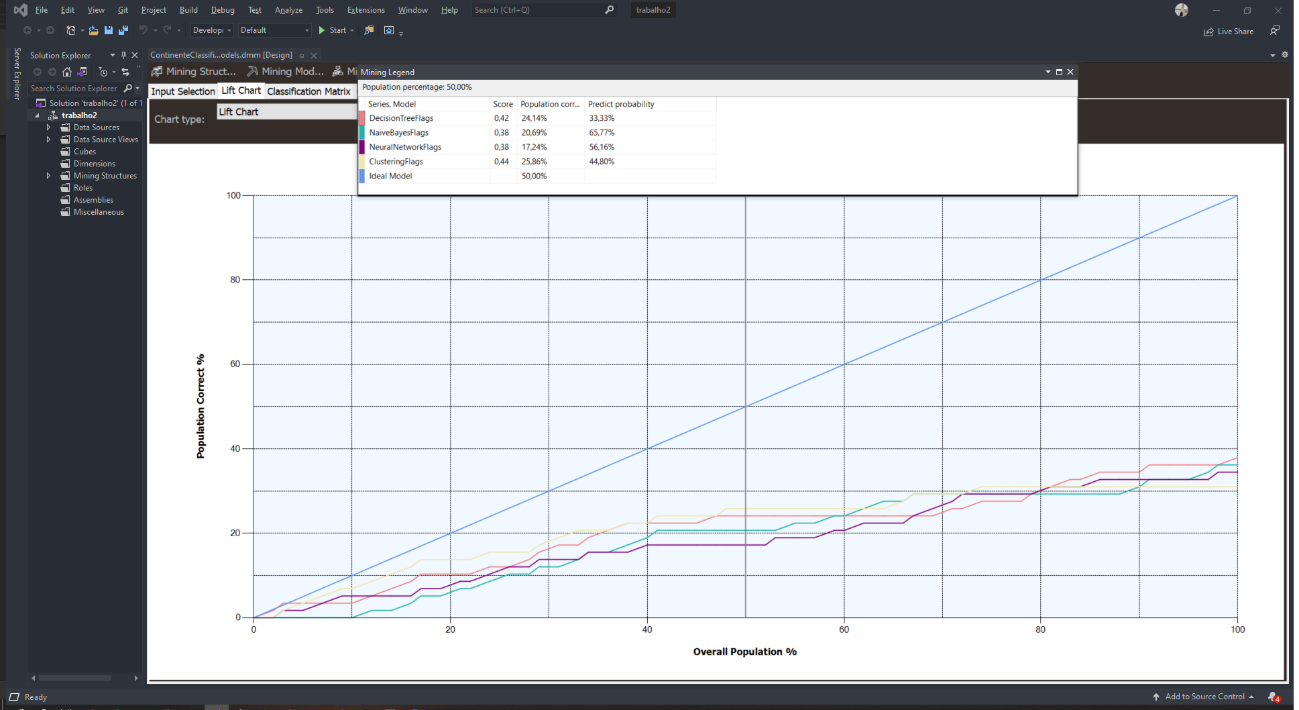


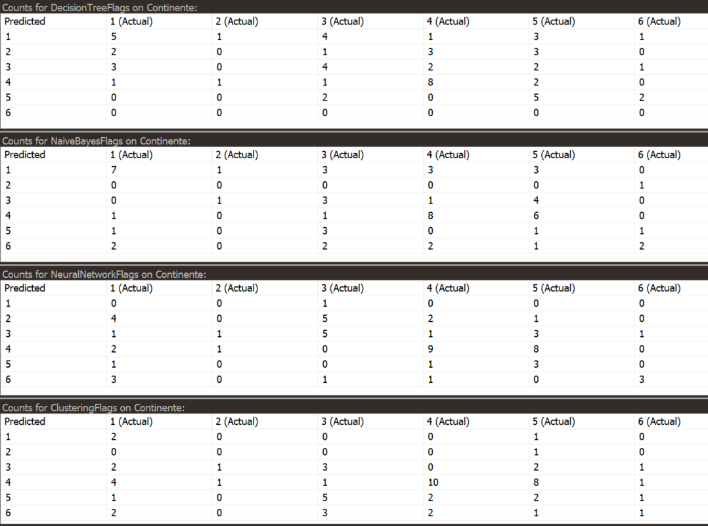






#### Conclusões

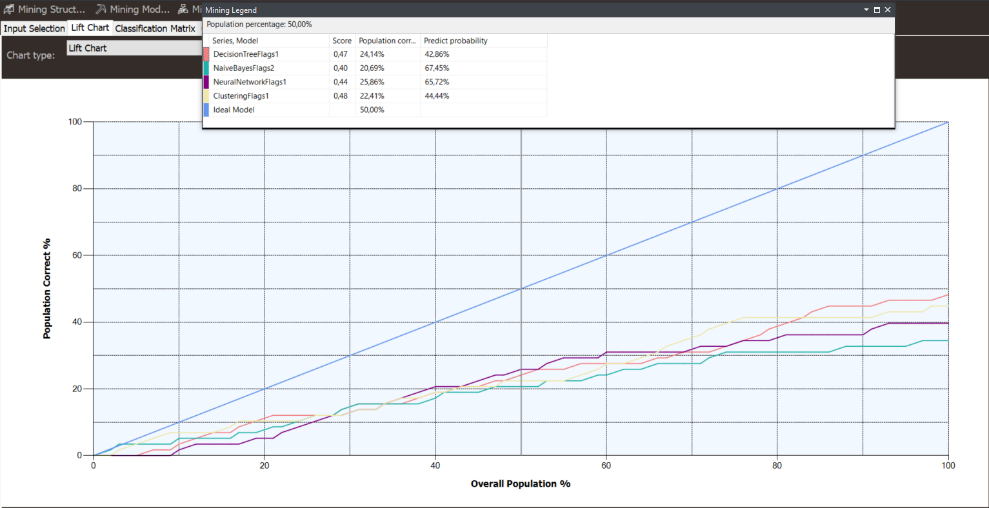
* Predicit: continente
* Key: Pais
* Inputs: Características das bandeiras

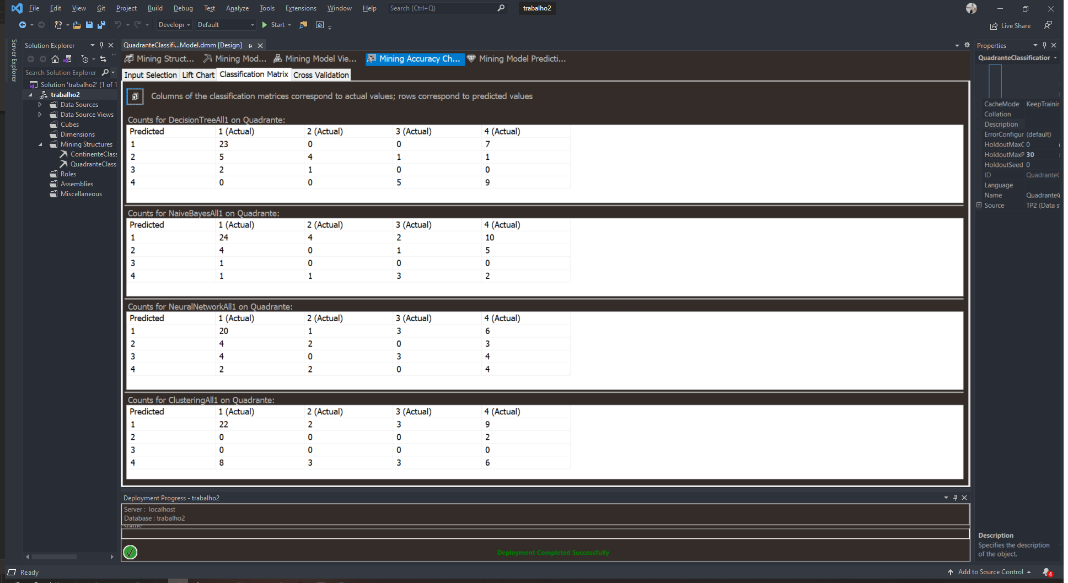


#### Cálculo das métricas

#### Conclusões

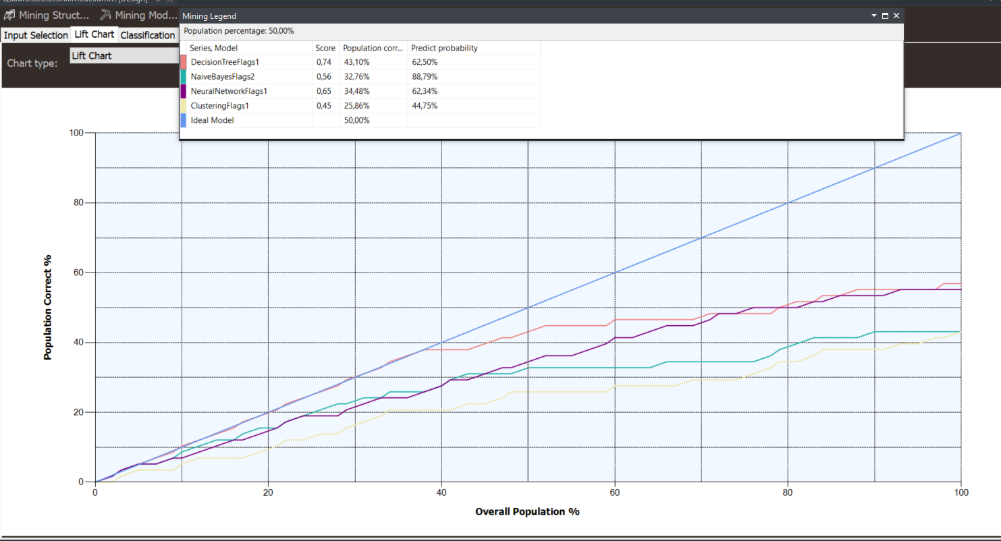
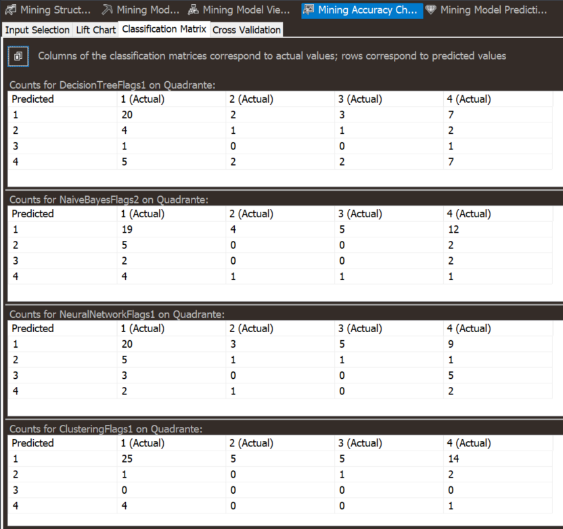
### Alvo: quadrante geográfico

* Predicit: quadrante
* Key: Pais
* Inputs: Todos atributos menos *língua*



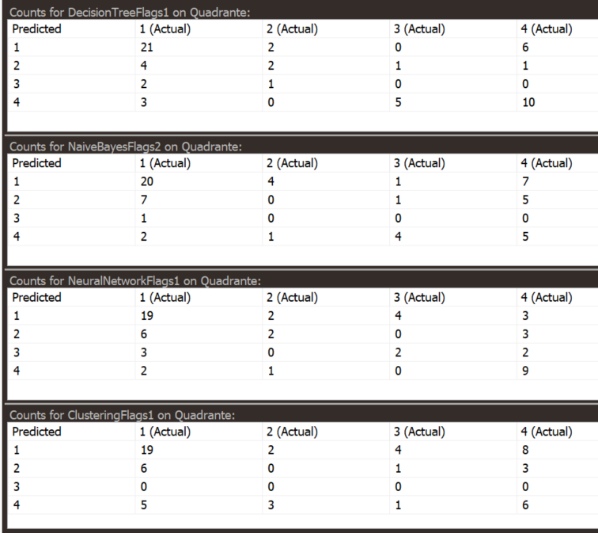
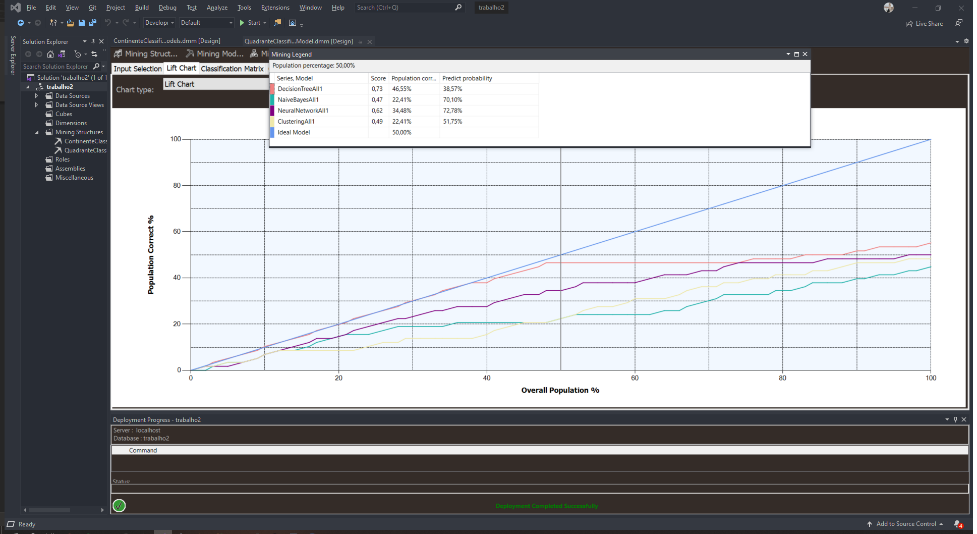
#### Cálculo das métricas

#### Conclusões

* Predicit: quadrante
* Key: Pais
* Inputs: Bandeiras apenas

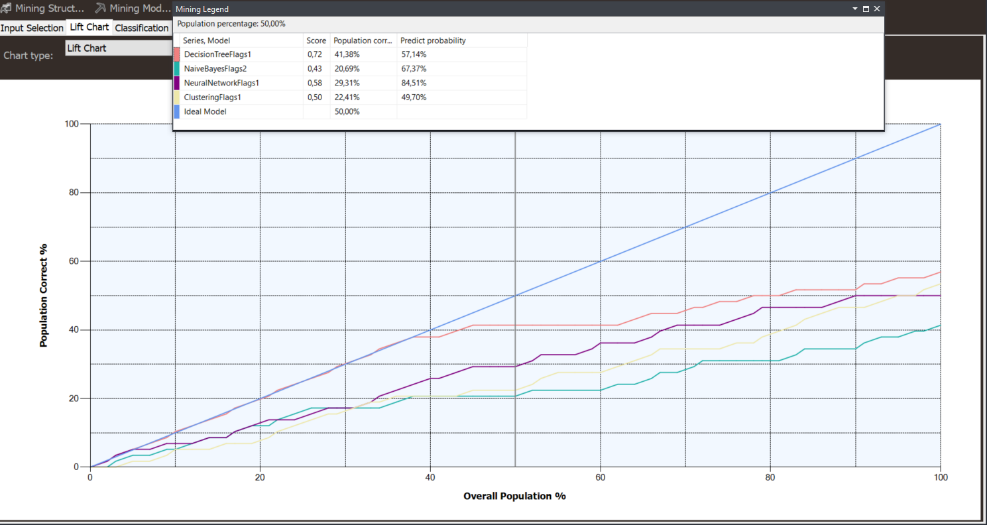
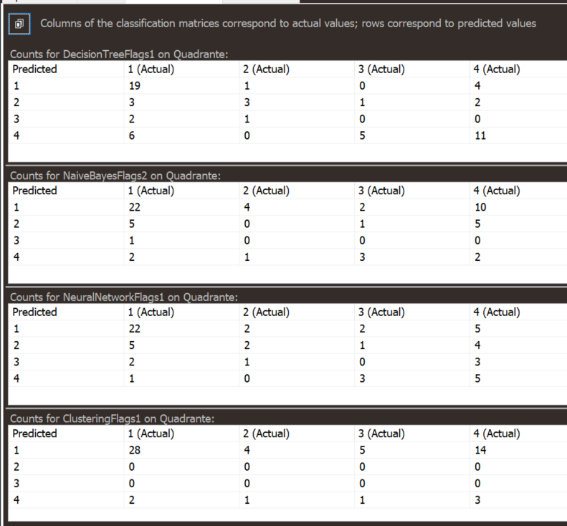
#### Cálculo das métricas

#### Conclusões

* Predicit: quadrante
* Key: Pais
* Inputs: Bandeiras + Continente

#### Cálculo das métricas

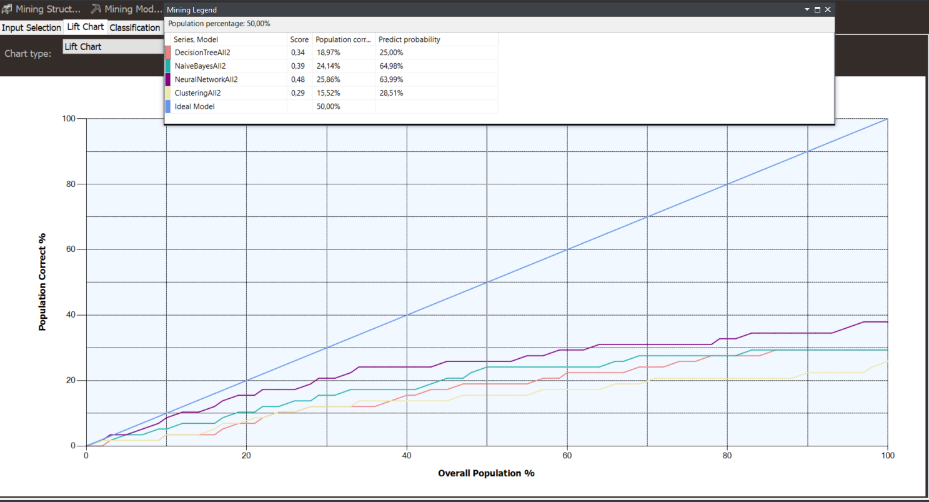
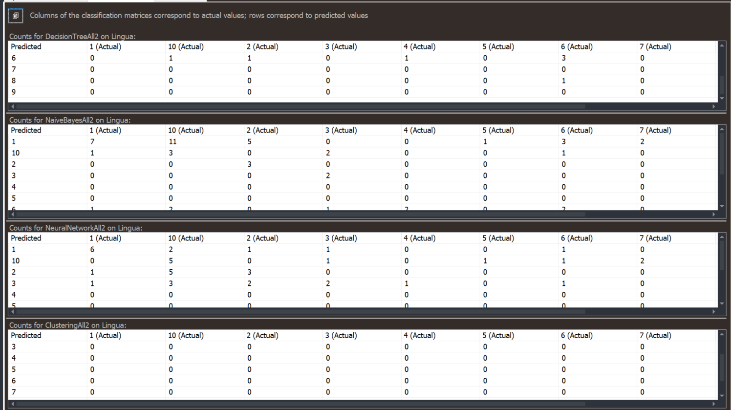
#### Conclusões

* Predicit: quadrante
* Key: Pais
* Inputs: Bandeiras + Continente + População

#### Cálculo das métricas

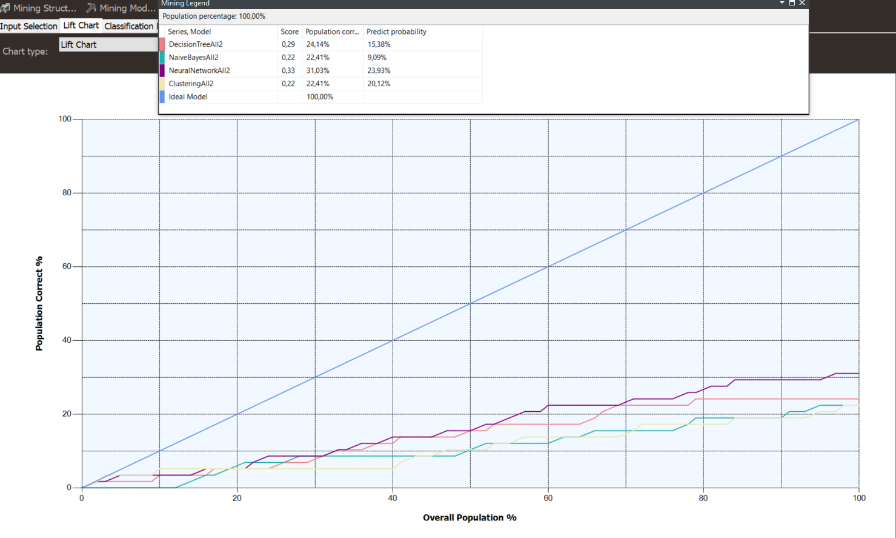
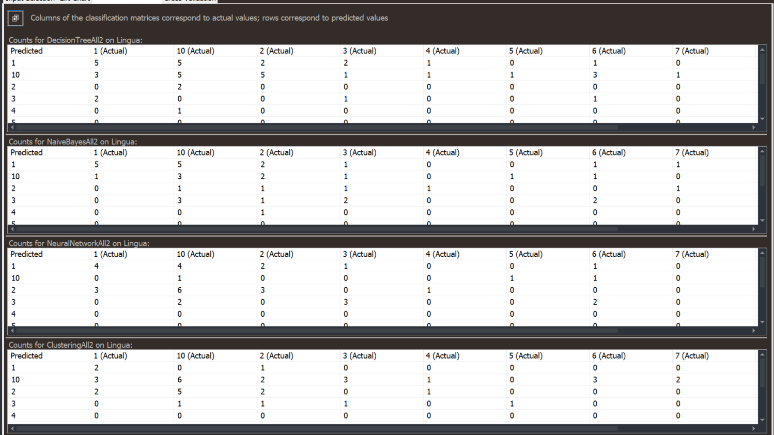
#### Conclusões

### Alvo: língua

* Predicit: lingua
* Key: Pais
* Inputs: All

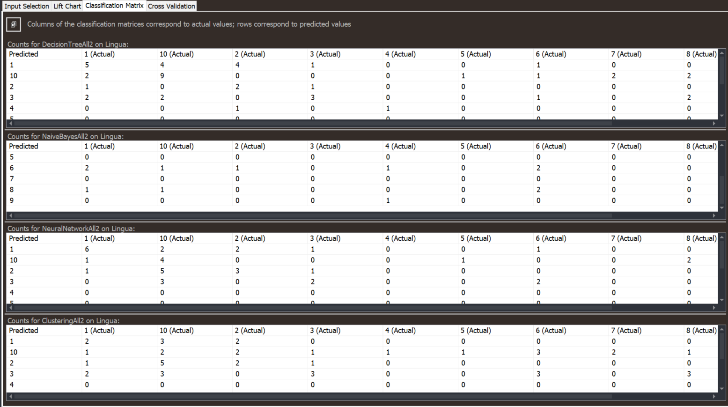
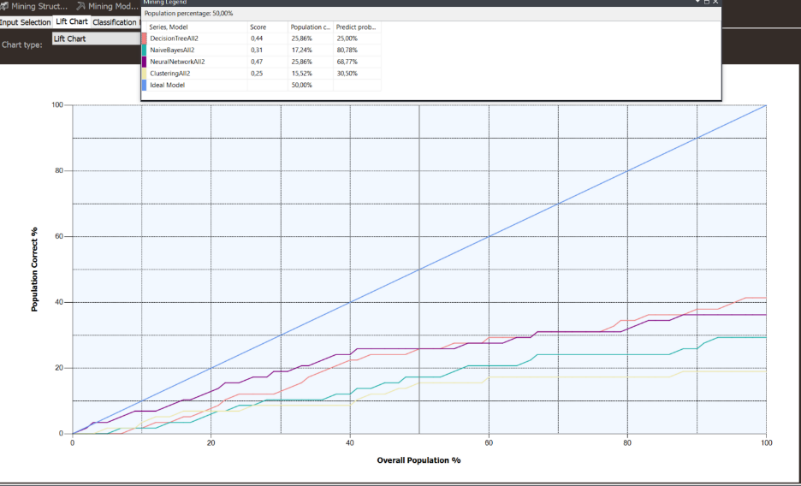
#### Cálculo das métricas

#### Conclusões

* Predicit: lingua
* Key: Pais
* Inputs: Bandeiras

#### Cálculo das métricas

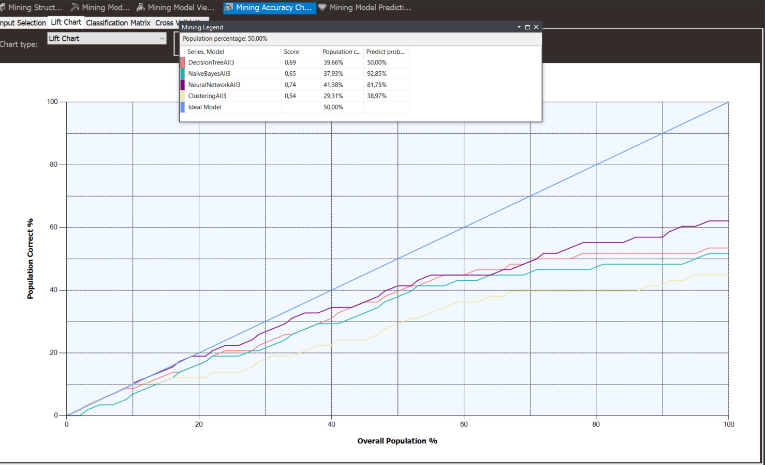
#### Conclusões

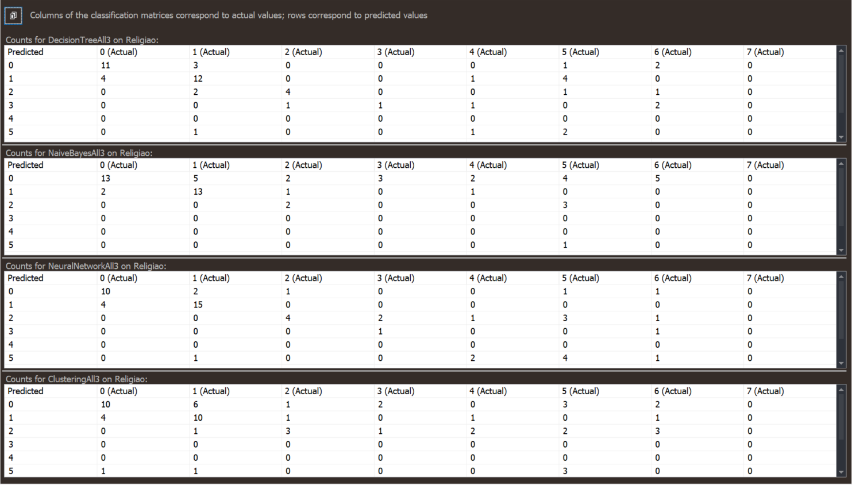
* Predicit: língua
* Key: Pais
* Inputs: Bandeiras + continente

#### Cálculo das métricas

#### Conclusões

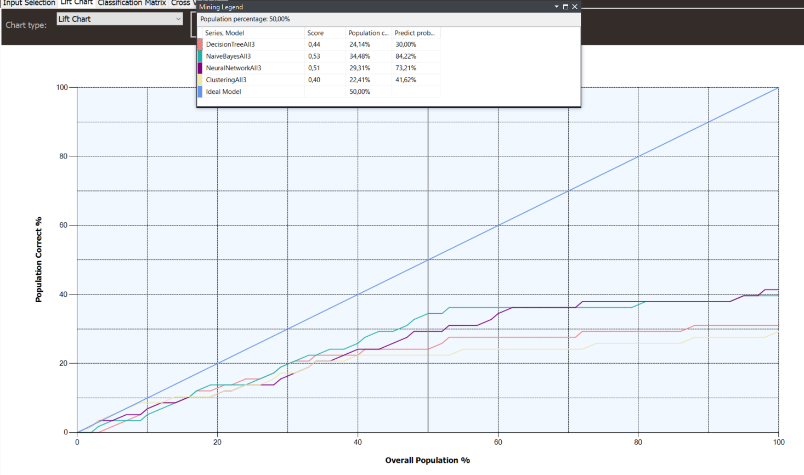
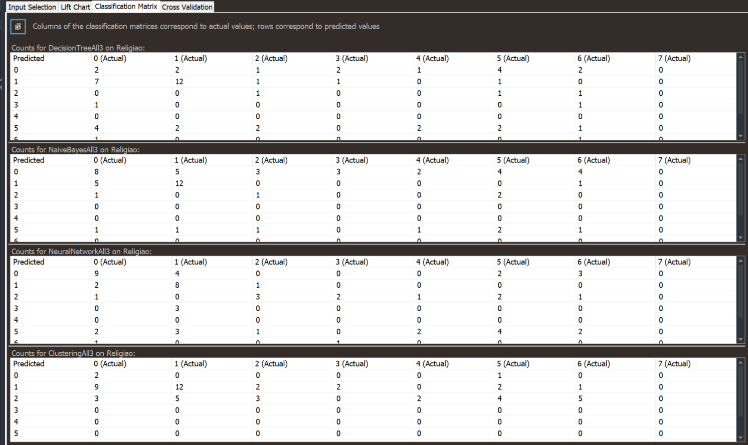
### Alvo: religião

* Predicit: religião
* Key: Pais
* Inputs: Bandeiras + todas



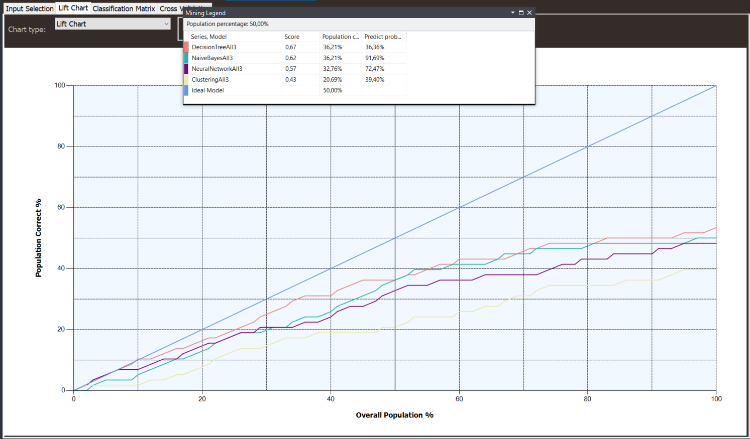
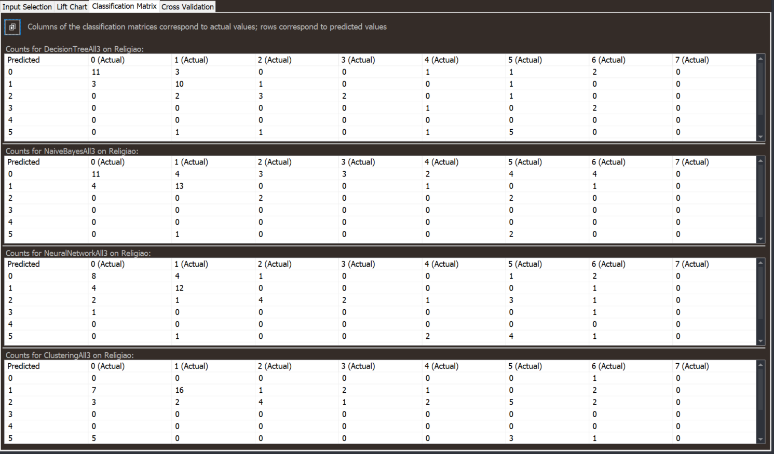
#### Cálculo das métricas

#### Conclusões

* Predicit: religião
* Key: Pais
* Inputs: Bandeiras + População

#### Cálculo das métricas

#### Conclusões

* Predicit: religião
* Key: Pais
* Inputs: Bandeiras + População + Continente

#### Cálculo das métricas

#### Conclusões

## Técnicas de clustering

Foi-nos pedido neste trabalho prático 2 para utilizar técnicas de clustering e interpretar os resultados que obtivemos. Assim, de seguida, apresentam-se os resultados obtidos e a respetiva análise.

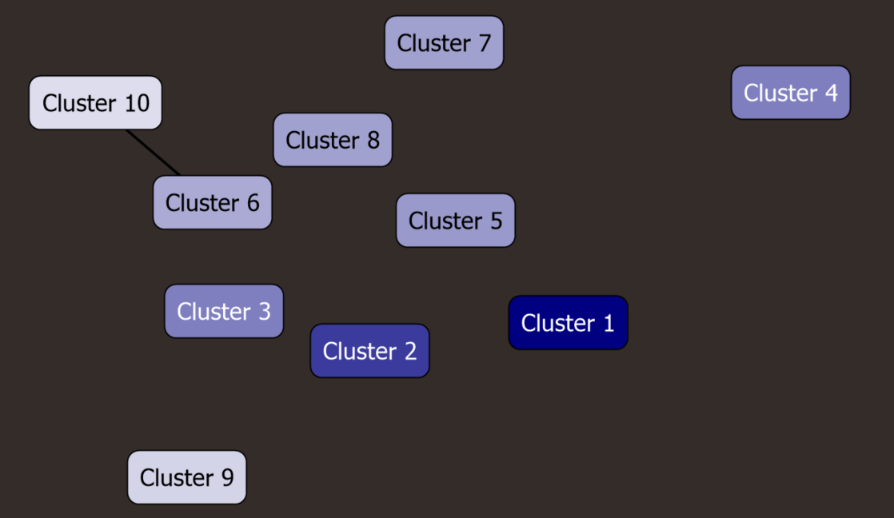
#### Clusters para alvo: continente

##### Cluster Diagram

#### 

Na imagem podemos observar os clusters que obtivemos, sendo que os que estão representados a mais escuro têm uma maior densidade e as linhas mais escuras representam as ligações mais fortes.

Nos 10 Clusters gerados e depois de reduzir a ligação entre os clusters para as relações mais fortes, podemos constatar pela imagem que se segue que a relação mais forte é entre o cluster 10 e 6.



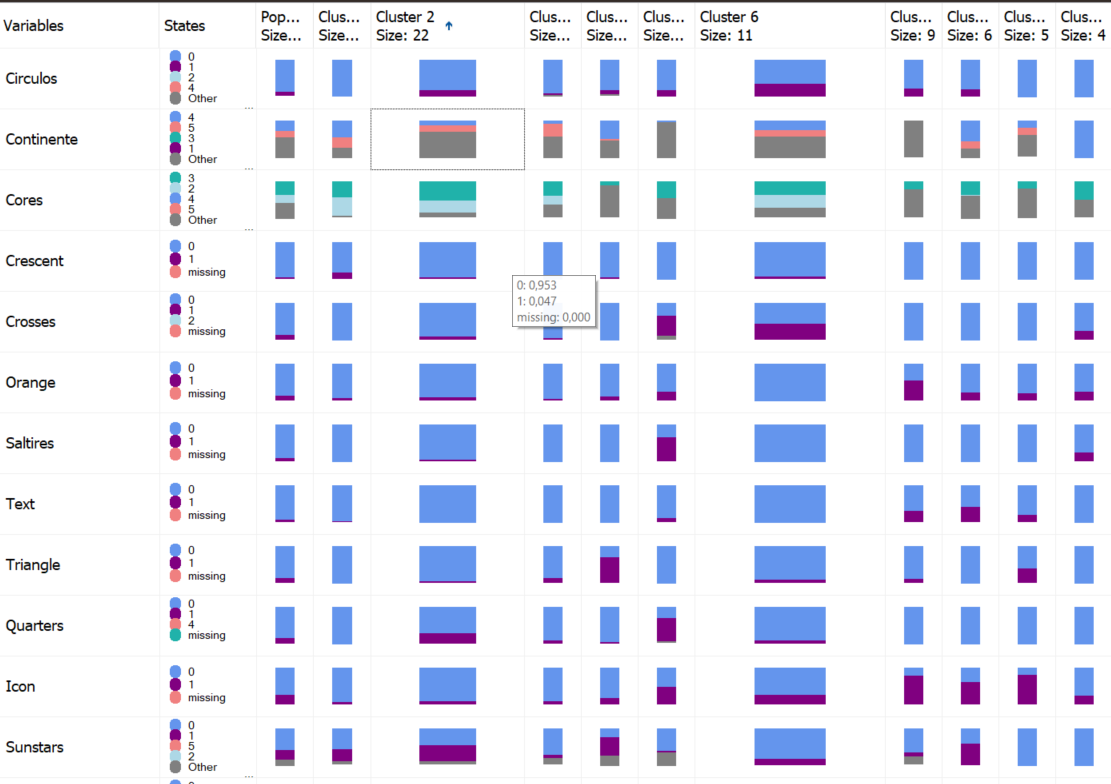
##### Cluster Profile

De seguida, apresentamos uma tabela que apresenta os vários perfis dos clusters, onde temos as variáveis da base de dados, os estados desse atributo, a população total, com 136 casos e, por fim, os respetivos clusters.

Assim, as conclusões que retiramos dos clusters 10 e 6 é o facto de ambos encontrarem o mesmo continente, isto pode ser corroborado pelo gráfico:

Continente no cluster 6: outro

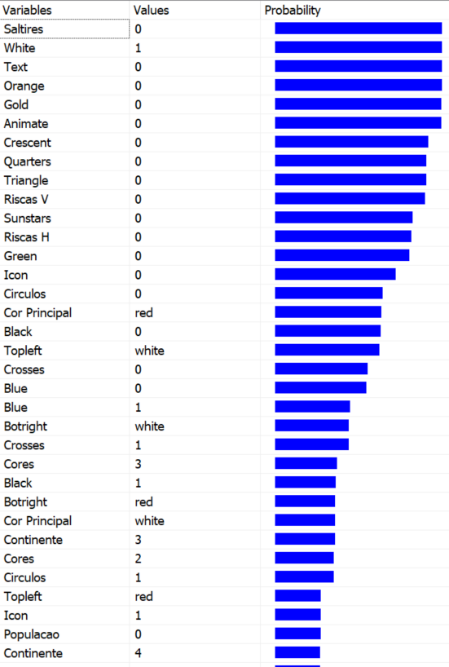
Continente no cluster 10: outro

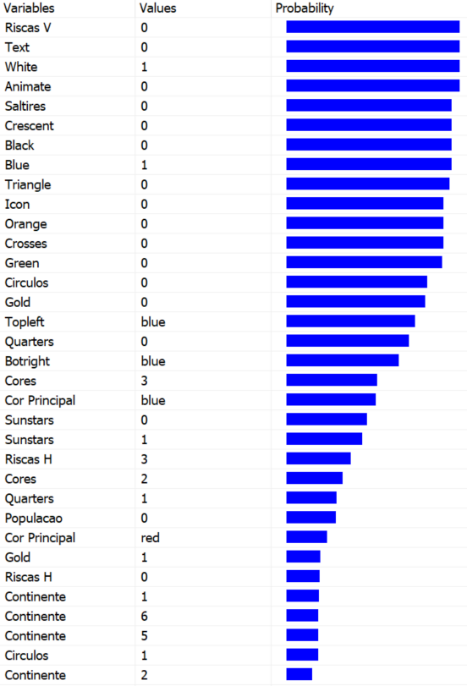
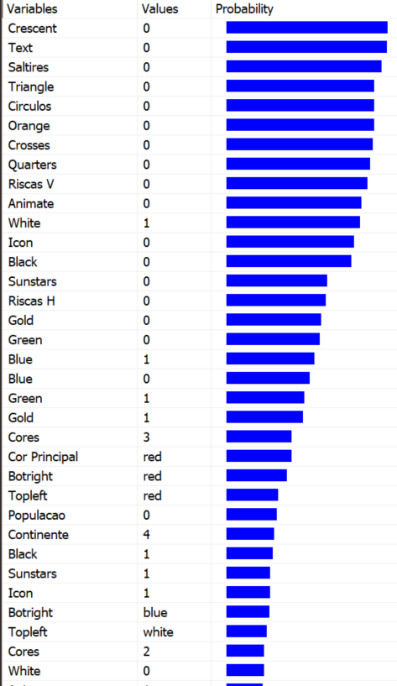


##### Características dos Clusters

Relativamente às características dos clusters, apresentamos, de seguida, as características com maior probabilidade para o total da população e depois para cada um dos clusters, de modo a percebermos as diferenças entre estes a nível de probabilidade, assim como as suas semelhanças.

No entanto, como se poderá constatar, a nível de interpretação não é tão explicita como o ponto anterior, apesar de nos dar os atributos com maiores probabilidades de uma forma direta.

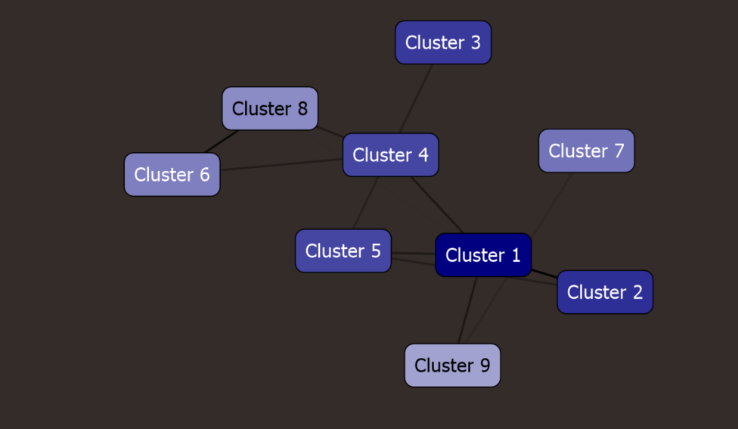
*Cluster 6 Cluster 10 População*

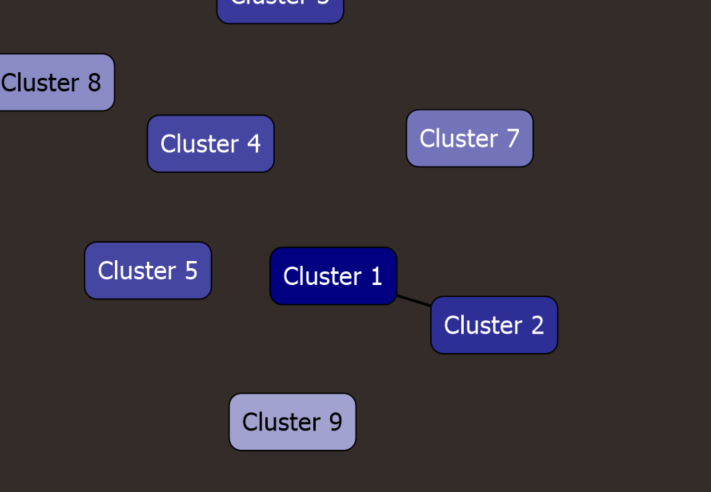


Observadas as características, podemos constatar que a totalidade da população só tem ligeiras diferenças em relação a diversos clusters, isto é, às suas variáveis.

#### Clusters para alvo: quadrante geográfico

##### Cluster Diagram

Na imagem podemos observar os clusters que obtivemos, sendo que os que estão representados a mais escuro têm uma maior densidade e as linhas mais escuras representam as ligações mais fortes.



Nos 10 Clusters gerados e depois de reduzir a ligação entre os clusters para as relações mais fortes, podemos constatar pela imagem que se segue que a relação mais forte é entre o cluster 1 e 2.

##### Cluster Profile

De seguida, apresentamos uma tabela que apresenta os vários perfis dos clusters, onde temos as variáveis da base de dados, os estados desse atributo, a população total, com 136 casos e, por fim, os respetivos clusters.

Assim, as conclusões que retiramos dos clusters 1 e 2 é o facto de ambos encontrarem o mesmo continente, isto pode ser corroborado pelo gráfico:

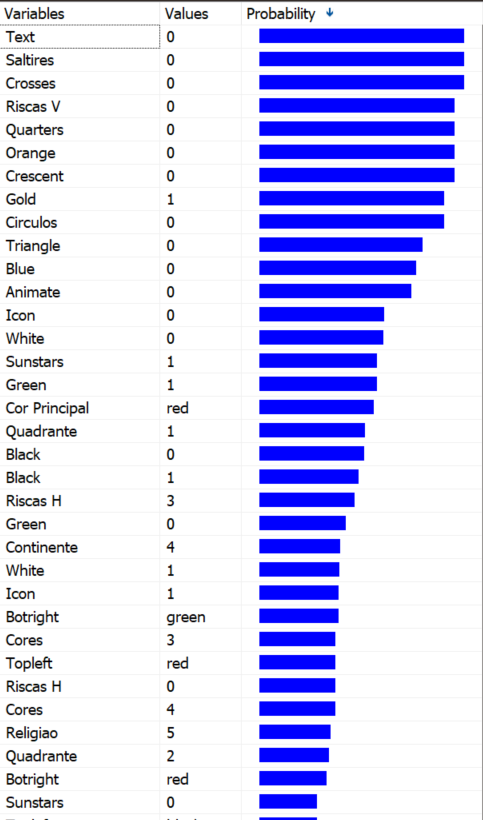
Quadrante no cluster 1: 1

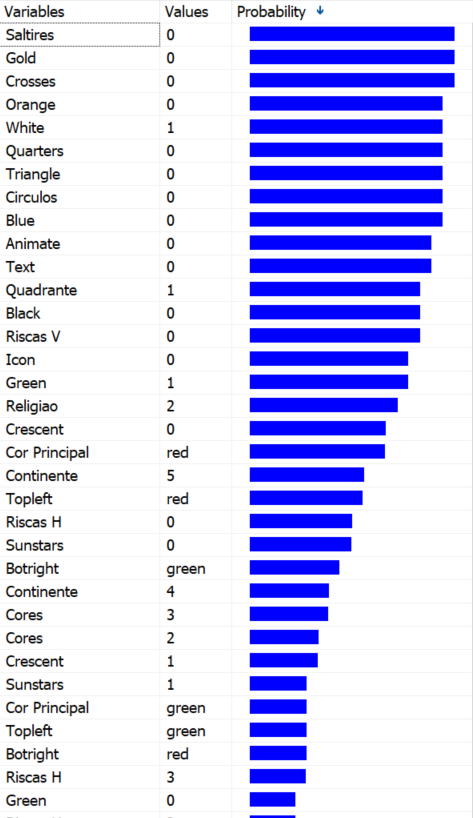
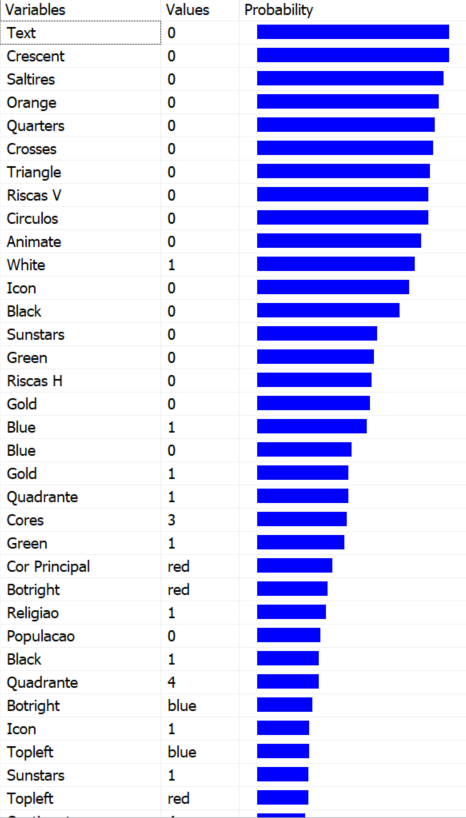
Quadrante no cluster 2: 1

##### Características dos Clusters

Relativamente às características dos clusters, apresentamos, de seguida, as características com maior probabilidade para o total da população e depois para cada um dos clusters, de modo a percebermos as diferenças entre estes a nível de probabilidade, assim como as suas semelhanças.

No entanto, como se poderá constatar, a nível de interpretação não é tão explicita como o ponto anterior, apesar de nos dar os atributos com maiores probabilidades de uma forma direta.

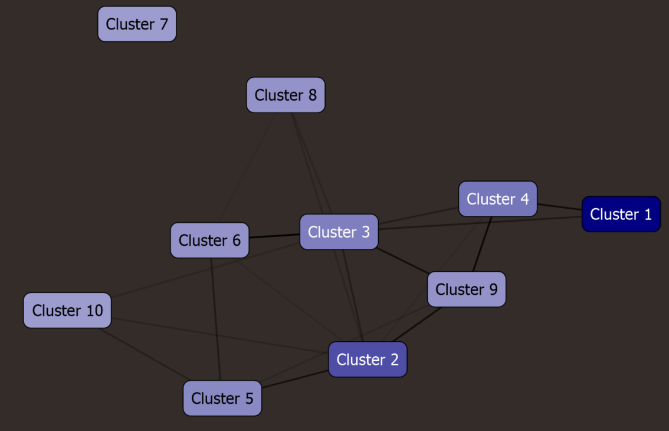
Cluster 1 Cluster 2 População

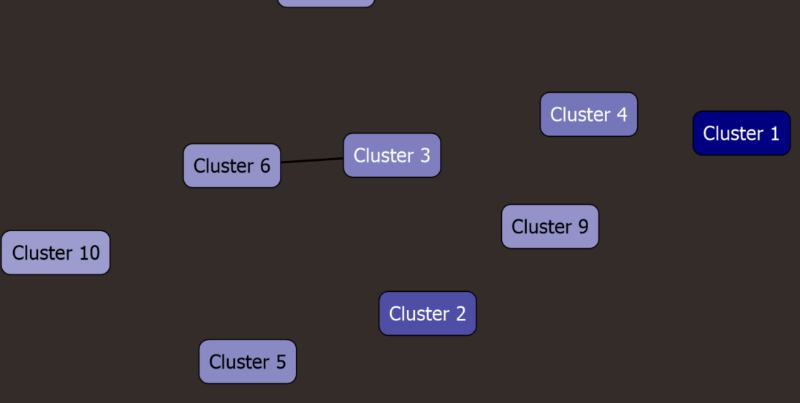


Observadas as características, podemos constatar que a totalidade da população só tem ligeiras diferenças em relação a diversos clusters, isto é, às suas variáveis.

#### Clusters para alvo: língua

##### Cluster Diagram

Na imagem podemos observar os clusters que obtivemos, sendo que os que estão representados a mais escuro têm uma maior densidade e as linhas mais escuras representam as ligações mais fortes.



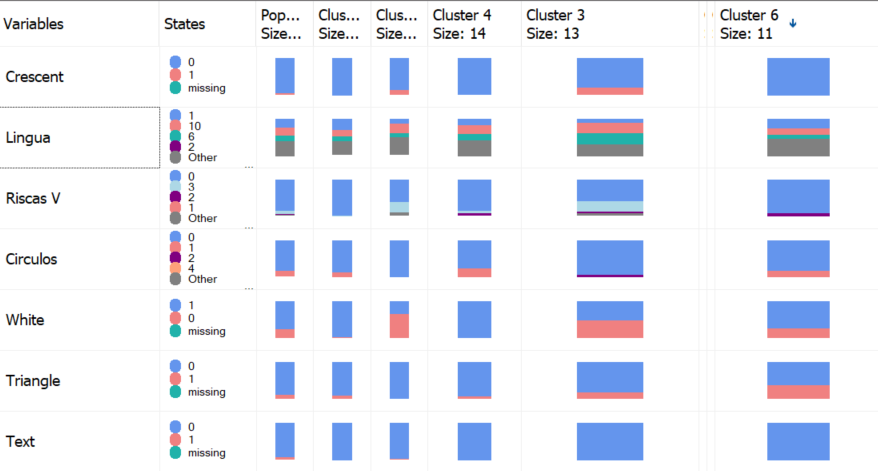
Nos 10 Clusters gerados e depois de reduzir a ligação entre os clusters para as relações mais fortes, podemos constatar pela imagem que se segue que a relação mais forte é entre o cluster 3 e 6.

##### Cluster Profile

De seguida, apresentamos uma tabela que apresenta os vários perfis dos clusters, onde temos as variáveis da base de dados, os estados desse atributo, a população total, com 136 casos e, por fim, os respetivos clusters.

Assim, as conclusões que retiramos dos clusters 3 e 6 é o facto de ambos encontrarem o mesmo continente, isto pode ser corroborado pelo gráfico:

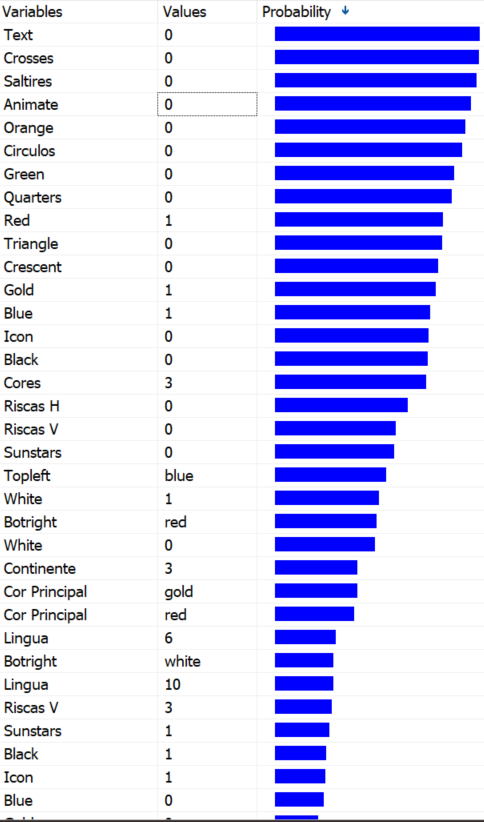
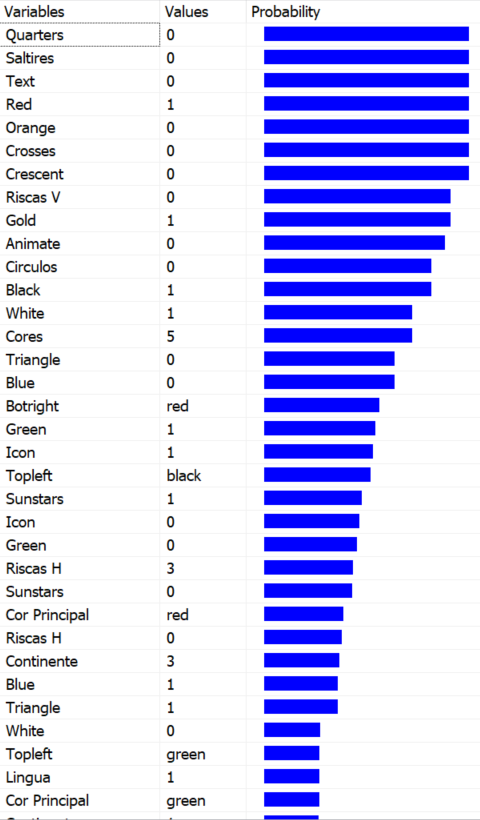
Língua no cluster 3: 6

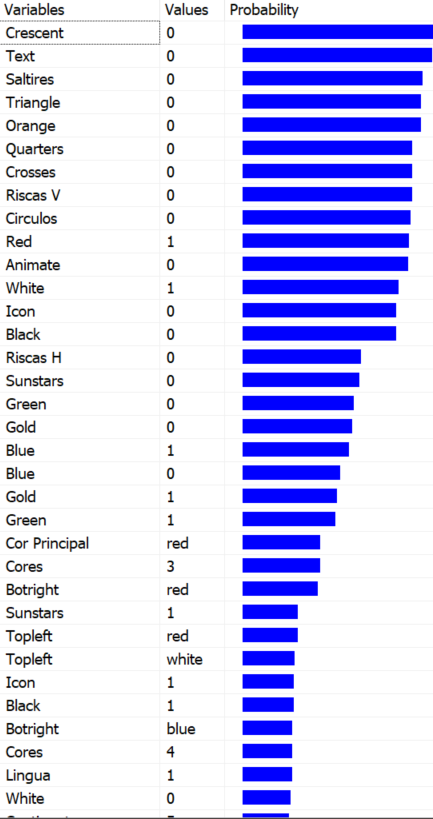
Língua no cluster 6: 1

##### Características dos Clusters

Relativamente às características dos clusters, apresentamos, de seguida, as características com maior probabilidade para o total da população e depois para cada um dos clusters, de modo a percebermos as diferenças entre estes a nível de probabilidade, assim como as suas semelhanças.

No entanto, como se poderá constatar, a nível de interpretação não é tão explicita como o ponto anterior, apesar de nos dar os atributos com maiores probabilidades de uma forma direta.

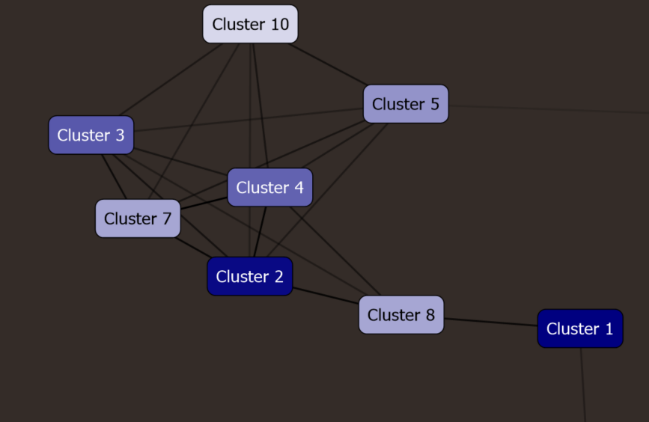
Cluster 3 Cluster 6 População

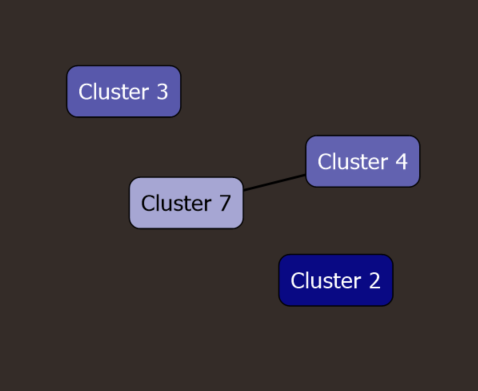


Observadas as características, podemos constatar que a totalidade da população só tem ligeiras diferenças em relação a diversos clusters, isto é, às suas variáveis.

#### Clusters para alvo: religião

##### Cluster Diagram

Na imagem podemos observar os clusters que obtivemos, sendo que os que estão representados a mais escuro têm uma maior densidade e as linhas mais escuras representam as ligações mais fortes.

Nos 10 Clusters gerados e depois de reduzir a ligação entre os clusters para as relações mais fortes, podemos constatar pela imagem que se segue que a relação mais forte é entre o cluster 4 e 7.

##### Cluster Profile

De seguida, apresentamos uma tabela que apresenta os vários perfis dos clusters, onde temos as variáveis da base de dados, os estados desse atributo, a população total, com 136 casos e, por fim, os respetivos clusters.

Assim, as conclusões que retiramos dos clusters 7 e 4 é o facto de ambos encontrarem o mesmo continente, isto pode ser corroborado pelo gráfico:

Religião no cluster 4: 5

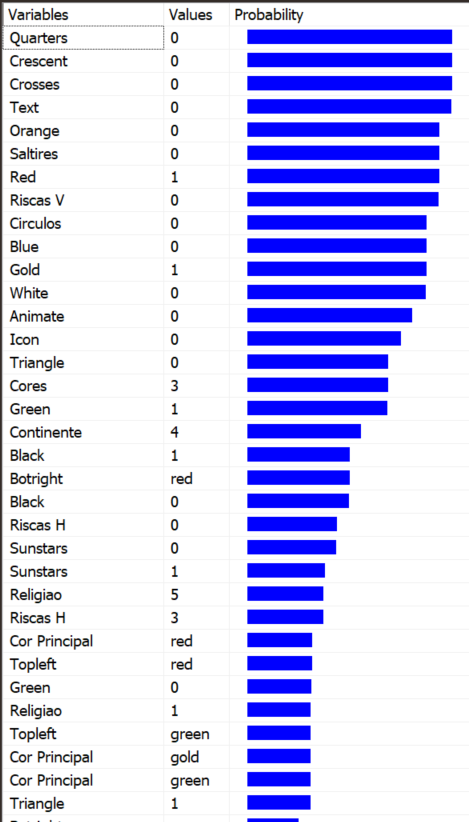
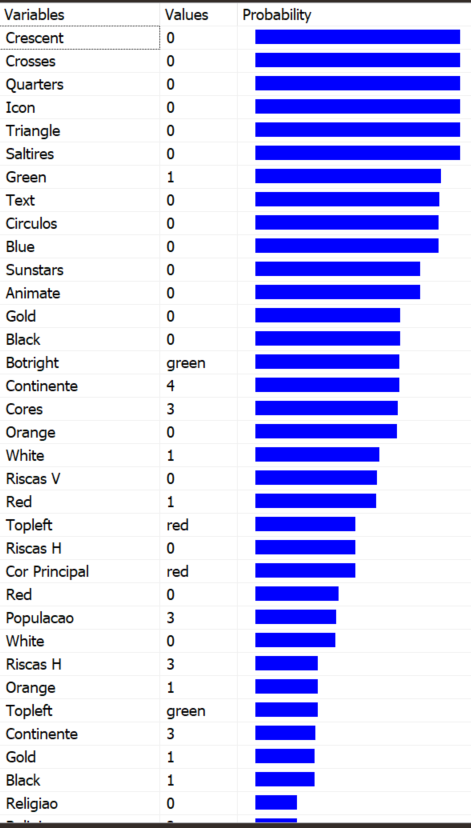
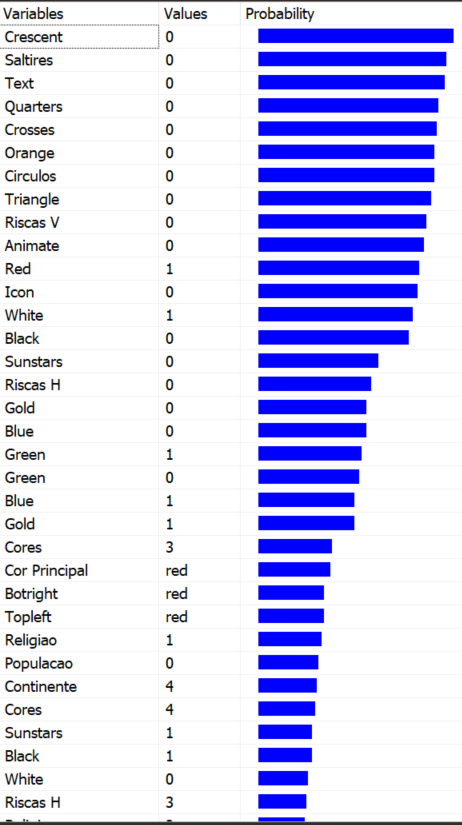
Religião no cluster 7: 5



##### Características dos Clusters

Relativamente às características dos clusters, apresentamos, de seguida, as características com maior probabilidade para o total da população e depois para cada um dos clusters, de modo a percebermos as diferenças entre estes a nível de probabilidade, assim como as suas semelhanças.

No entanto, como se poderá constatar, a nível de interpretação não é tão explicita como o ponto anterior, apesar de nos dar os atributos com maiores probabilidades de uma forma direta.

Cluster 4 Cluster 7 População

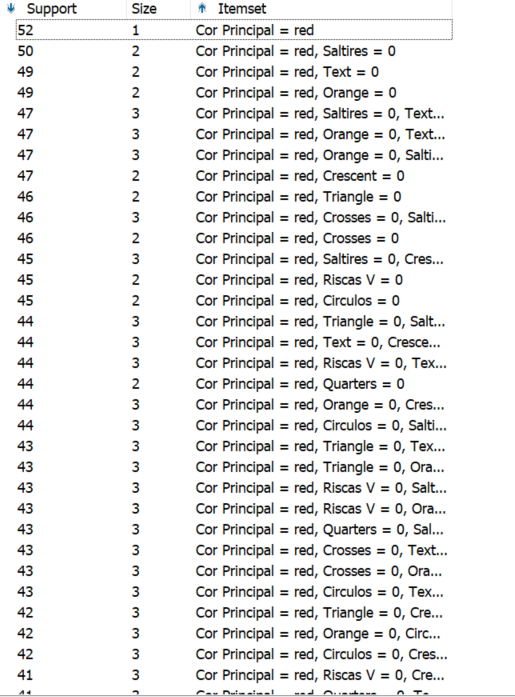
Observadas as características, podemos constatar que a totalidade da população só tem ligeiras diferenças em relação a diversos clusters, isto é, às suas variáveis.

## Regras de associação

#### Alvo: Continente

As regras de associação são um método baseado em regras para descobrir relações interessantes entre variáveis. Na secção das Rules, decidimos ver quais eram as probabilidades para cada saída do predict “Continente”, de forma a ter noção de qual seria a probabilidade de ocorrência deste.

## 

Como podemos observar pela imagem, ao mesmo nível de importância e probabilidade temos que o continente acertado é o 4 sempre sendo que um atributo importante nesta associação ao continente 4 é a população.

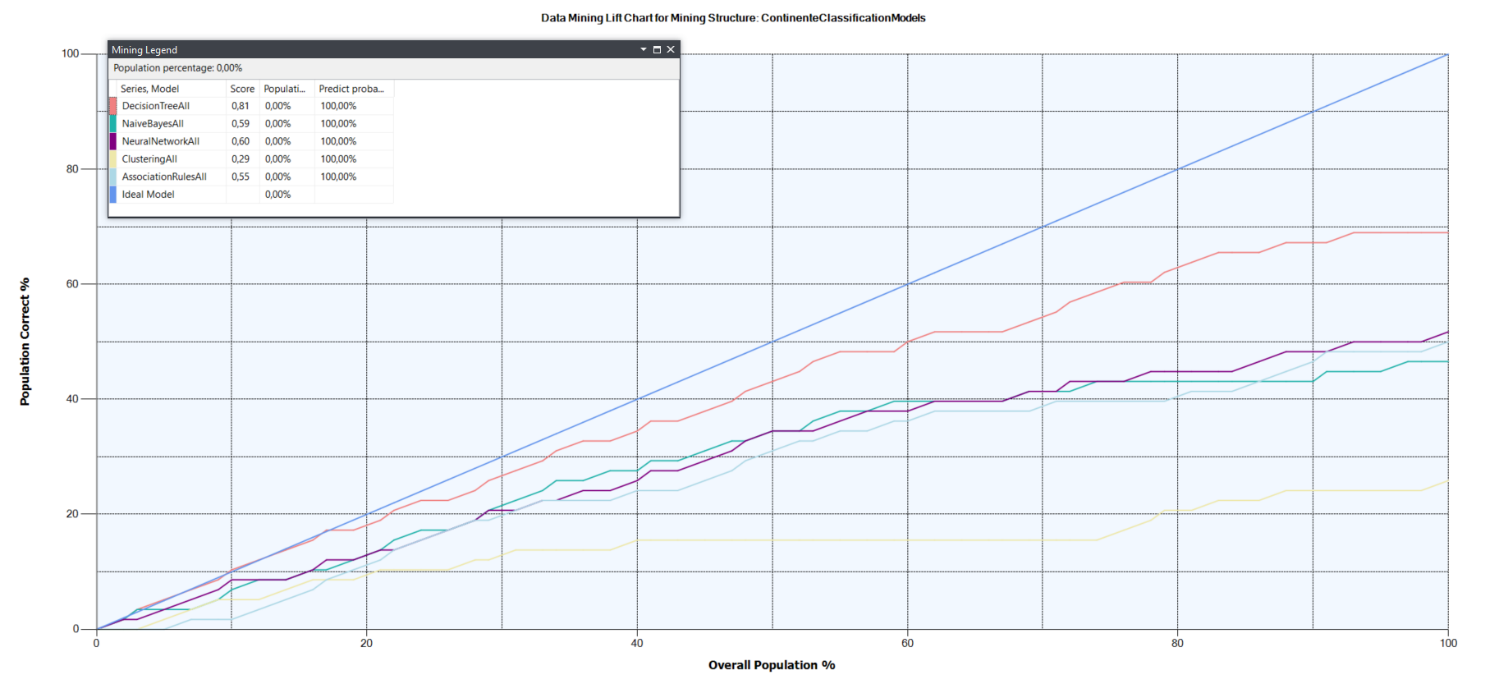
##### ItemSet (Melhor Resultado)

Na imagem seguinte vemos o

melhor resultado que obtivemos

para o Itemset e podemos concluir

que a cor que mais aparece é a cor red.



Assim, o melhor resultado para o alvo continente, considerando todos os inputs das bandeiras foi obtido com o Decision Tree ( 81,0%), seguido do Neural Nework (60%). O Association Rules teve um valor de 55%

## Regras de associação

#### Alvo: Quadrante geográfico

As regras de associação são um método baseado em regras para descobrir relações interessantes entre variáveis. Na secção das Rules, decidimos ver quais eram as probabilidades para cada saída do predict “Quadrante Geográfico”, de forma a ter noção de qual seria a probabilidade de ocorrência deste.

## 

Como podemos observar pela imagem, ao mesmo nível de importância e probabilidade temos que o continente acertado é o 4 sempre sendo que um atributo importante nesta associação ao continente 4 é a população.

##### 

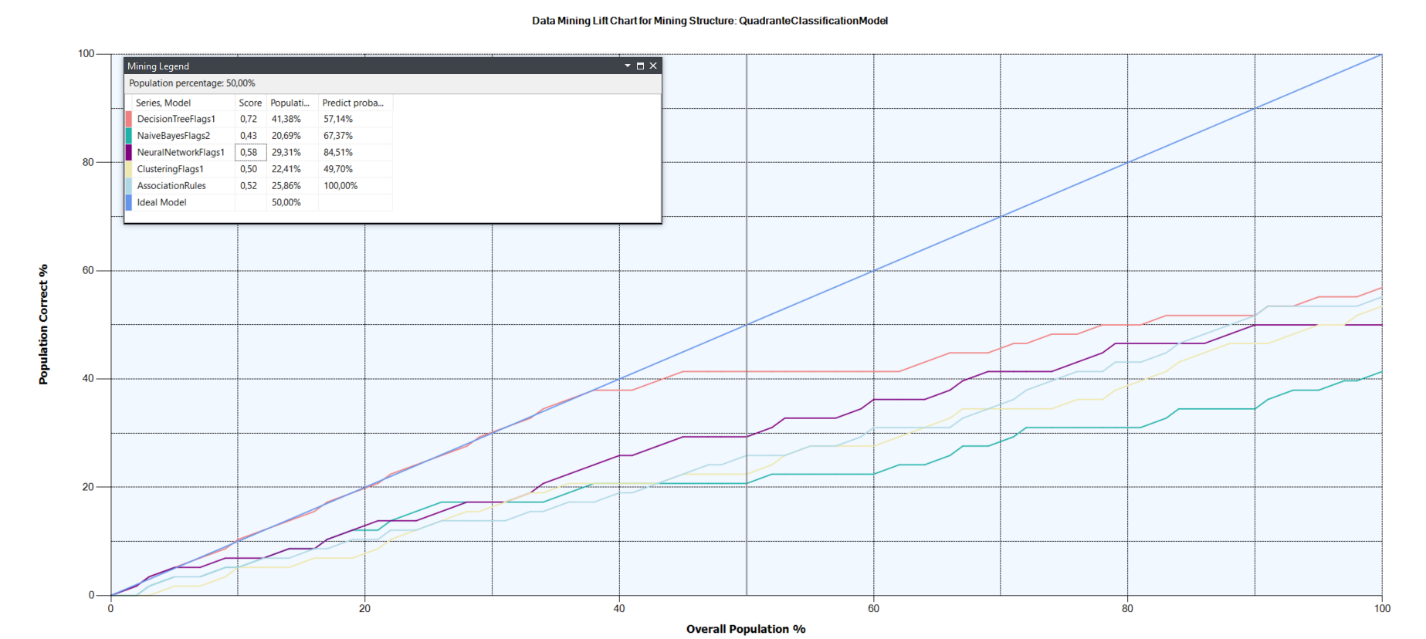
##### ItemSet (Melhor Resultado)

Na imagem seguinte vemos o

melhor resultado que obtivemos

para o Itemset e podemos concluir

que a cor que mais aparece é a cor red.



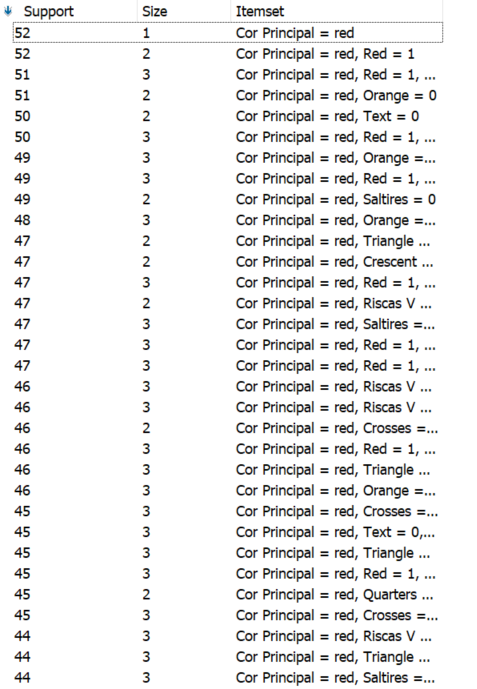
Assim, o melhor resultado para o alvo quadrante geográfico, considerando todos os inputs das bandeiras foi obtido com o Decision Tree (72,0%), seguido do Neural Nework (58%). O Association Rules teve um valor de 52%

## Regras de associação

#### Alvo: Língua

As regras de associação são um método baseado em regras para descobrir relações interessantes entre variáveis. Na secção das Rules, decidimos ver quais eram as probabilidades para cada saída do predict “Língua”, de forma a ter noção de qual seria a probabilidade de ocorrência deste.

## 

Como podemos observar pela imagem, ao mesmo nível de importância e probabilidade temos que o continente acertado é o 4 sempre sendo que um atributo importante nesta associação ao continente 4 é a população.

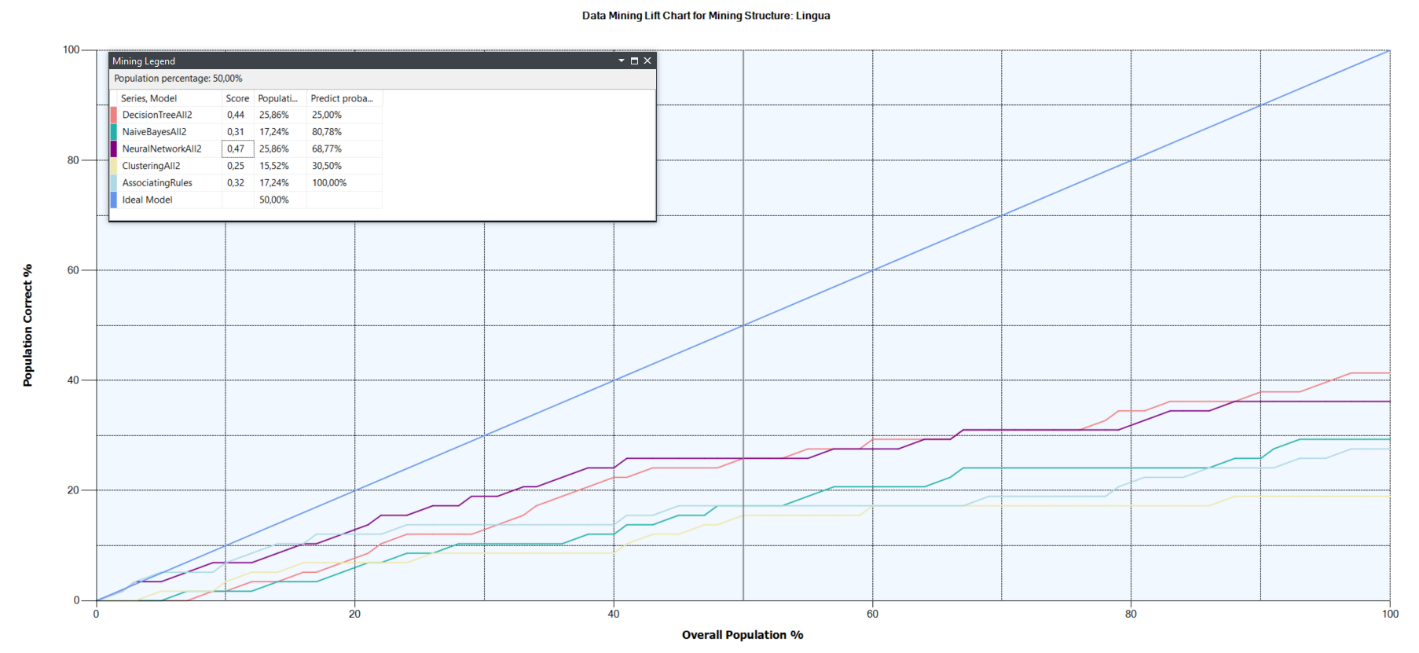
##### ItemSet (Melhor Resultado)

Na imagem seguinte vemos o

melhor resultado que obtivemos

para o Itemset e podemos concluir

que a cor que mais aparece é a cor red.



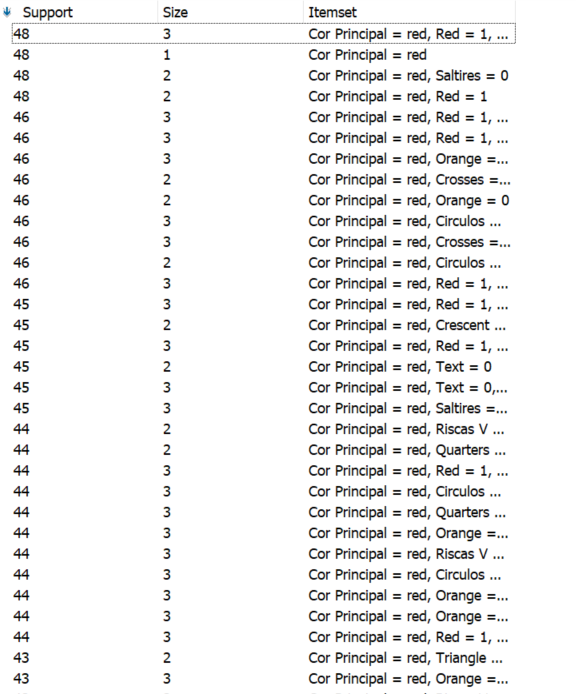
Assim, o melhor resultado para o alvo continente, considerando todos os inputs das bandeiras foi obtido com o NeuralNetwork (47,0%), seguido do DecisionTreel (44%). O Association Rules teve um valor de 32%

## Regras de associação

#### Alvo: Religião

As regras de associação são um método baseado em regras para descobrir relações interessantes entre variáveis. Na secção das Rules, decidimos ver quais eram as probabilidades para cada saída do predict “Religião”, de forma a ter noção de qual seria a probabilidade de ocorrência deste.

## 

Como podemos observar pela imagem, ao mesmo nível de importância e probabilidade temos que a religião acertado é a 1 sempre sendo que um atributo importante nesta associação ao continente 4 é a população.

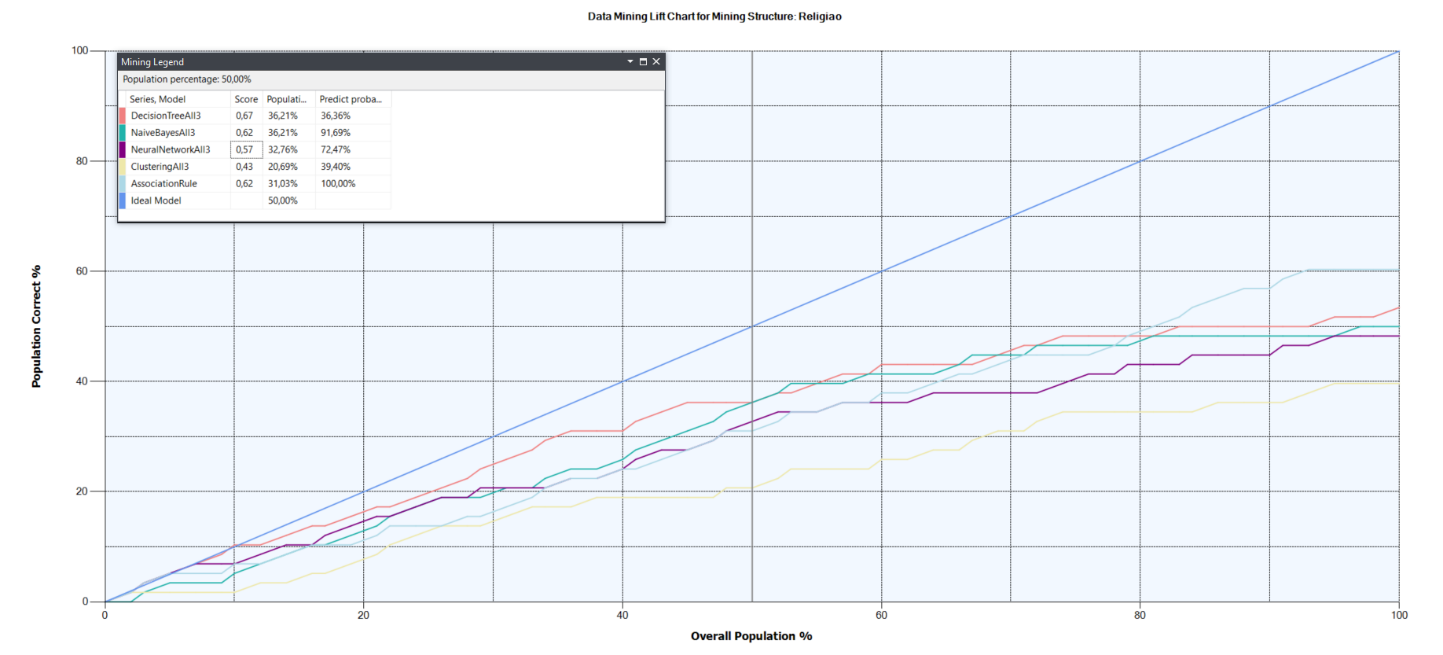
##### ItemSet (Melhor Resultado)

Na imagem seguinte vemos o

melhor resultado que obtivemos

para o Itemset e podemos concluir

que a cor que mais aparece é a cor red.



Assim, o melhor resultado para o alvo religião, considerando todos os inputs das bandeiras foi obtido com o Decision Tree ( 67,0%), seguido do NaiveBayes (62%) e do Association Rules que também teve um valor de 62%.

## Conclusão