EXTRAÇÃO DE CONHECIMENTO E DADOS PARA A GESTÃO

PÓS GRADUAÇÃO EM GESTÃO DOS SISTEMAS DE INFORMAÇÃO EMPRESARIAIS

P.PORTO

INSTITUTO SUPERIOR DE CONTABILIDADE E ADMINISTRAÇÃO DO PORTO

GRUPO1:

BRUNA TOMÉ · 12190209 JOÃO PEREIRA · 12190211 RUI RIBEIRO · 6940007

Índice da Apresentação

1. Descrição do Domínio

2. Preparação e limpeza dos dados

3. Análise Descritiva

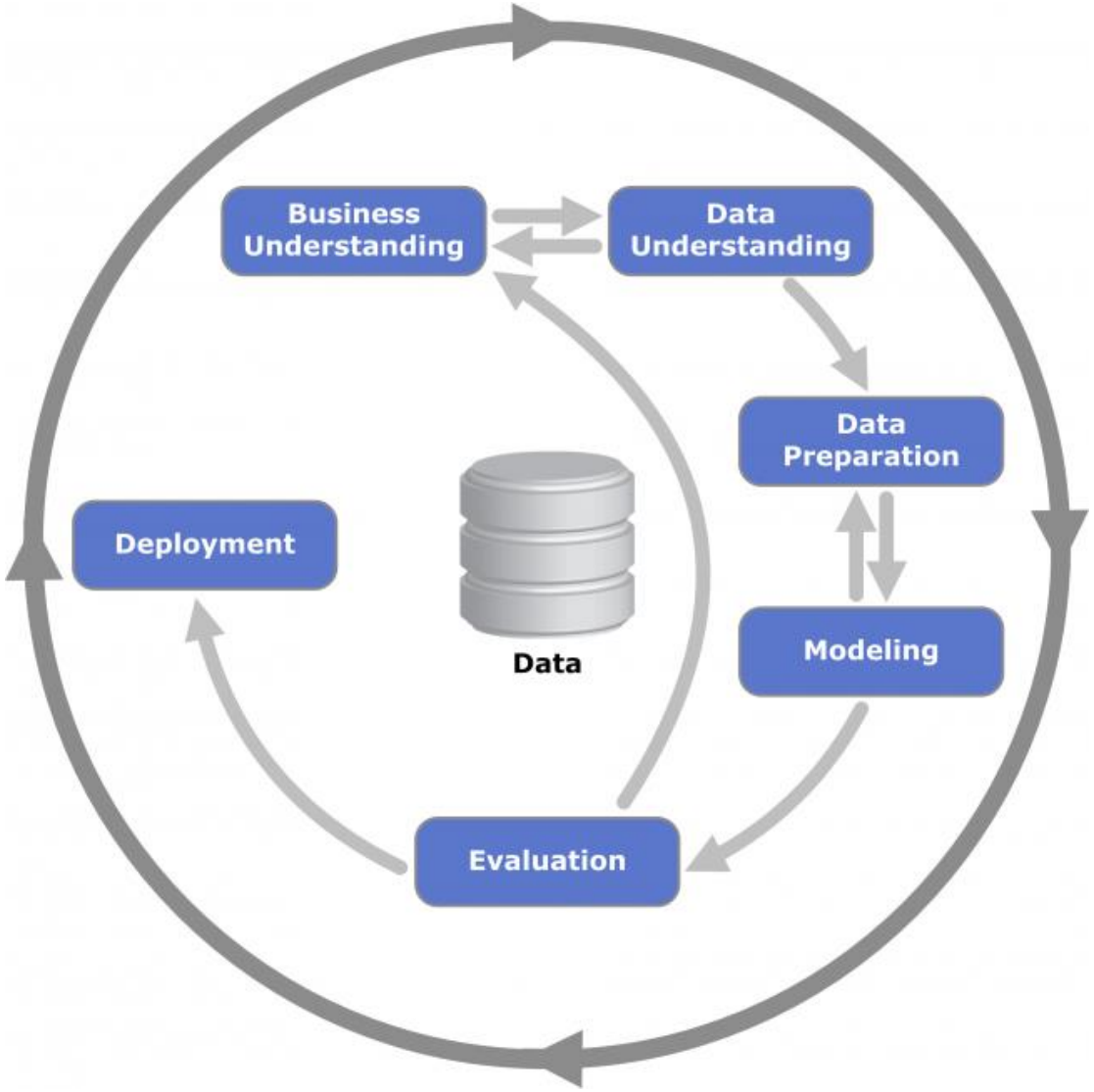
4. Análise Preditiva

5. Conclusões, limitações e trabalho futuro

Metodologia Utilizada:

CRISP-DM (Cross-Industry Standard Process for Data Mining)

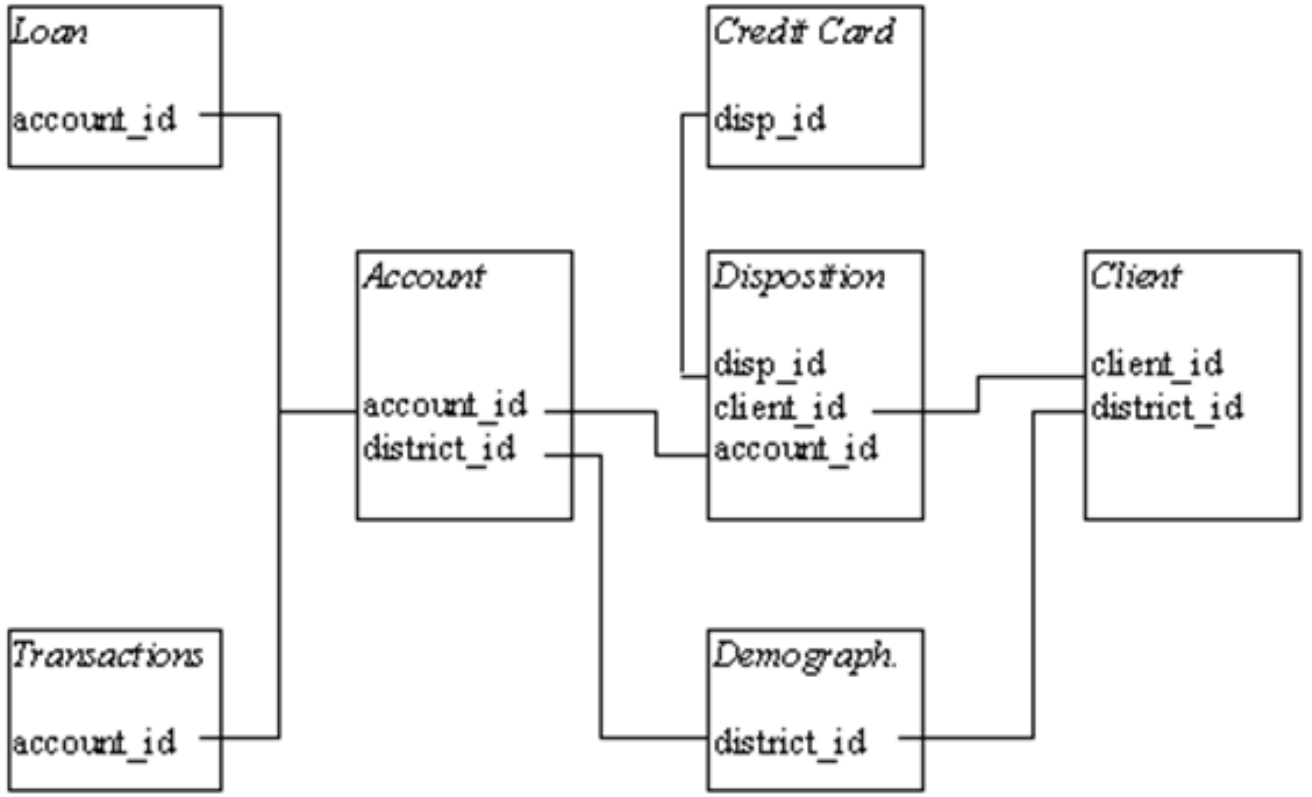
- Conhecimento do Negócio
- Conhecimento dos Dados
- Preparação dos Dados
- Modelação
- Avaliação
- Implementação



1. DESCRIÇÃO DO DOMÍNIO

Descrição do Domínio

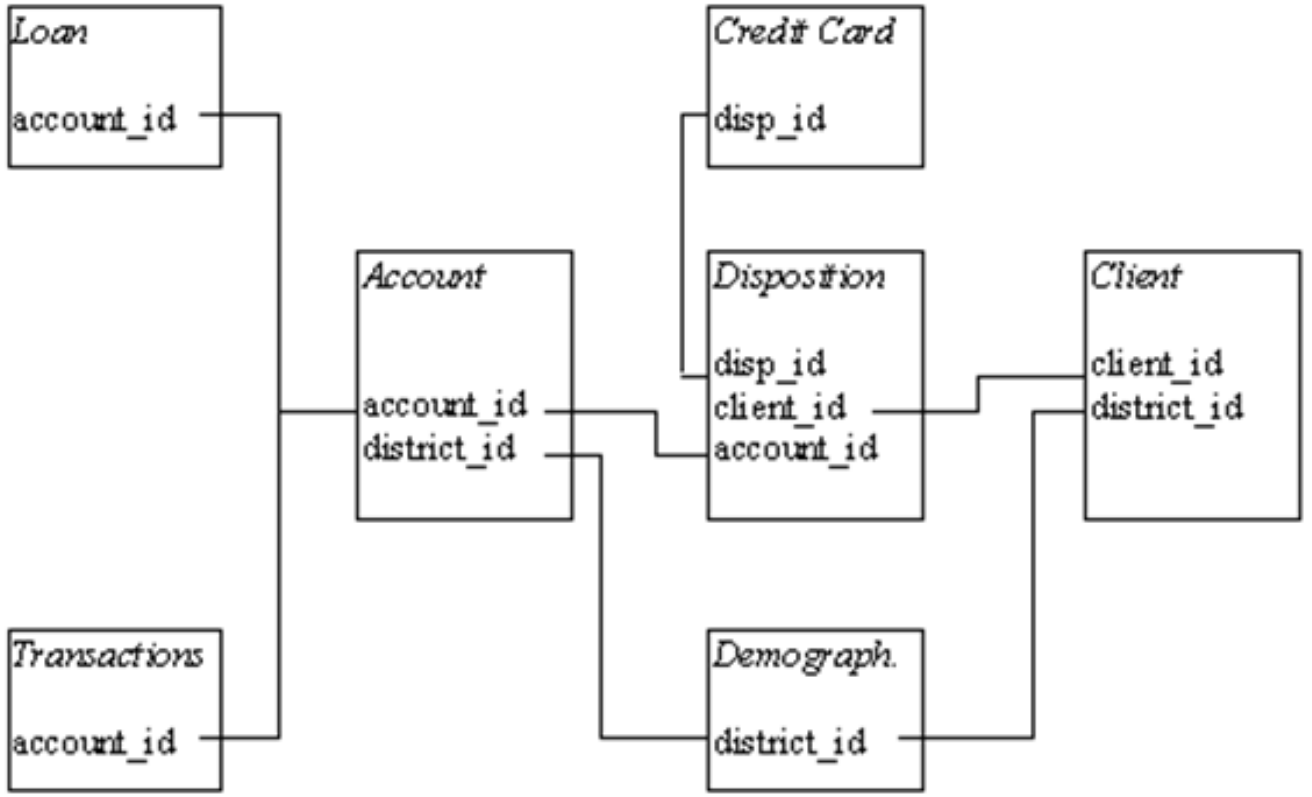
- O estudo do caso teve como principal objetivo a análise de dados de empréstimos de um Banco Checo, entre os anos de 1993 e 1996.
- Inclui informação relativa a clientes, dados geográficos, contas bancárias, empréstimos, cartões de crédito e movimentos bancários.
- Os dados encontram-se divididos entre dados de treino (análise histórica) e de teste (análise preditiva).



Descrição do Domínio

Os dados são compostos por:

- 4500 contas bancárias
- 5369 clientes
- 77 registos relativos a localizações geográficas
- 328 registos relativos a histórico de empréstimos
- 396,685 de registos relativos a histórico de transações



2. PREPARAÇÃO E LIMPEZA DOS DADOS

Preparação e limpeza dos dados

- Eliminar dados em falta
- Conversão de dados binomiais em dados numéricos
- Definir *roles*, *Ids* e *Labels*
- Converter campos numéricos em campos formato data
- Criação do atributo género do cliente
- Calcular a idade e data nascimento dos clientes

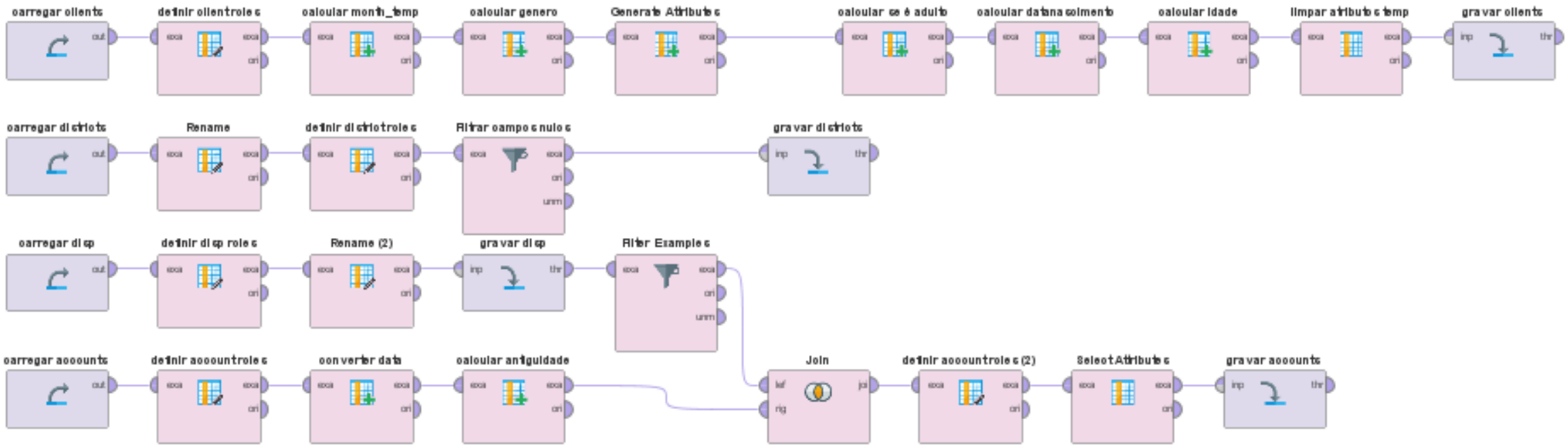
Preparação e limpeza dos dados

- Calcular se os clientes são adultos
- Calcular a antiguidade da conta (em dias)
- Determinar o proprietário da conta
- Converter os movimentos para débitos ou créditos
- Calcular o saldo médio, mínimo e máximo da conta
- Calcular o saldo à data do pedido de empréstimo

Preparação e limpeza dos dados

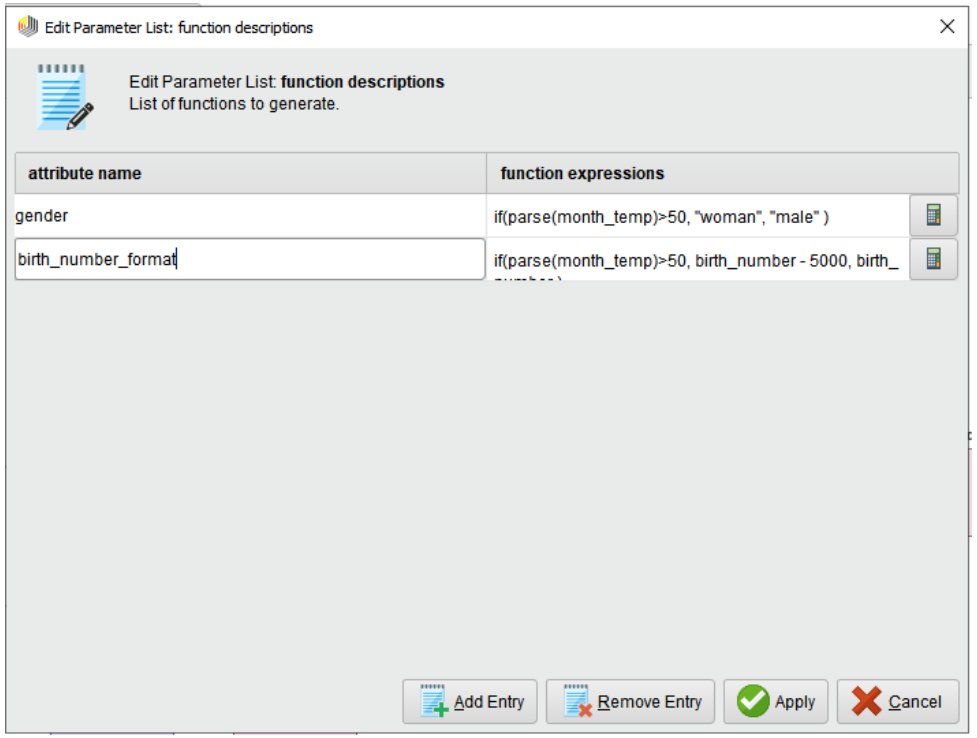
- Calcular totais agregados por tipologia de movimento (entradas de bancos, saídas para cartão de crédito, etc)
- Calcular o total de movimentos associados à conta bancária
- Determinar se existe cartão de crédito associado à conta bancária
- Converter registos de localização em falta, através de regressão linear.

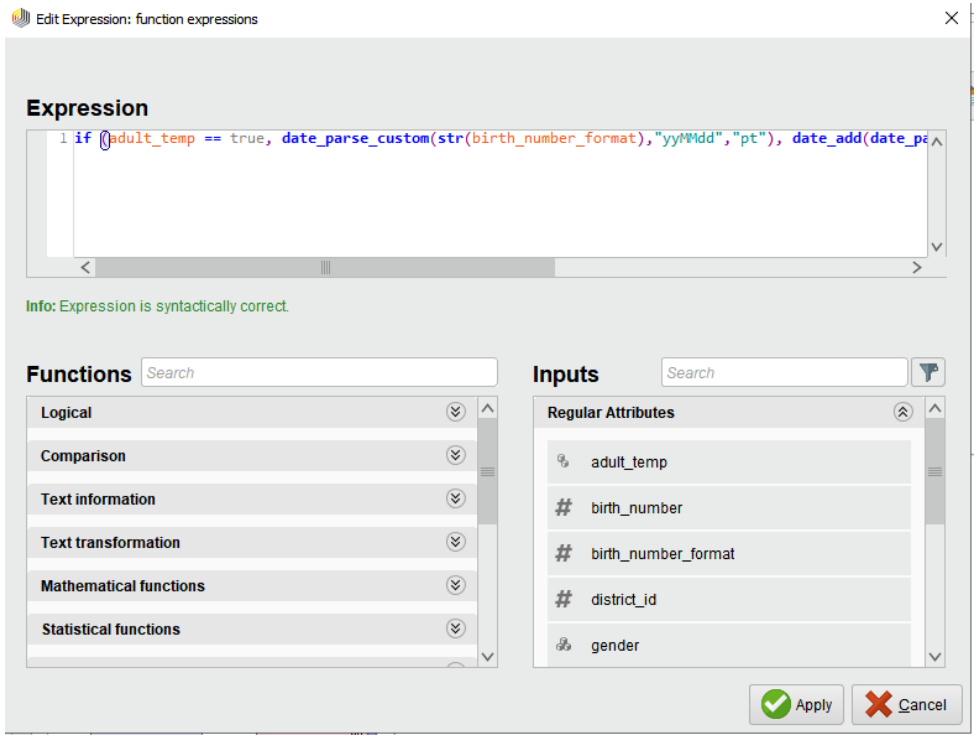
Preparação e limpeza dos dados



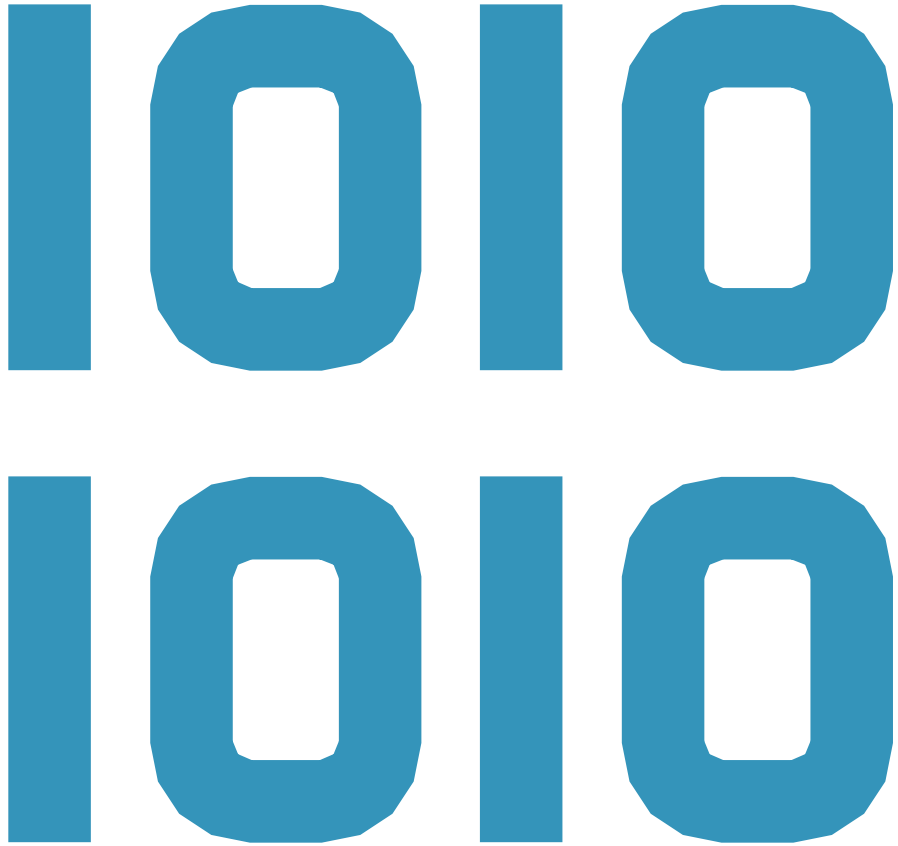
Preparação e limpeza dos dados

Exemplos de cálculo do género e da idade





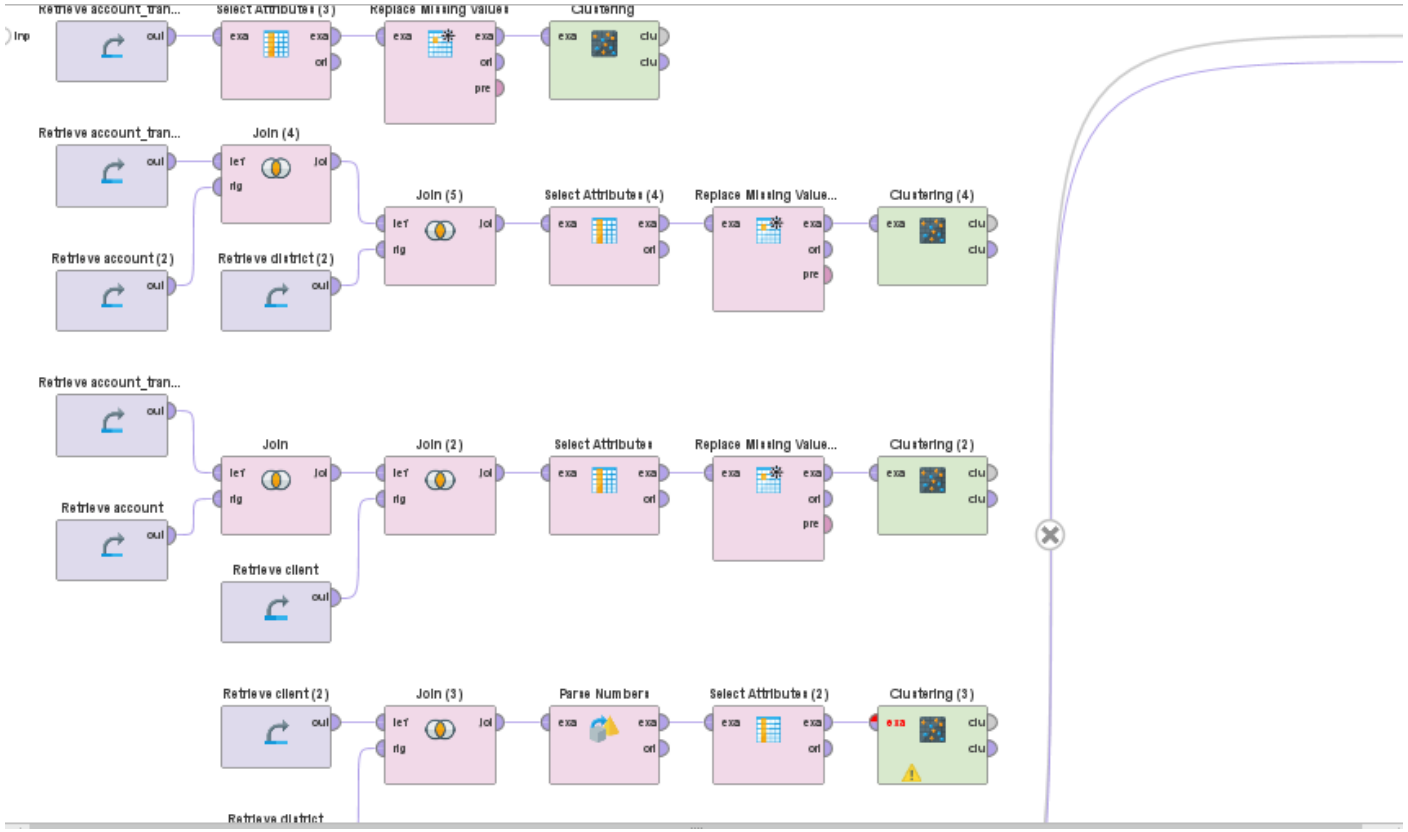
3. ANÁLISE DESCRITIVA



Análise Descritiva

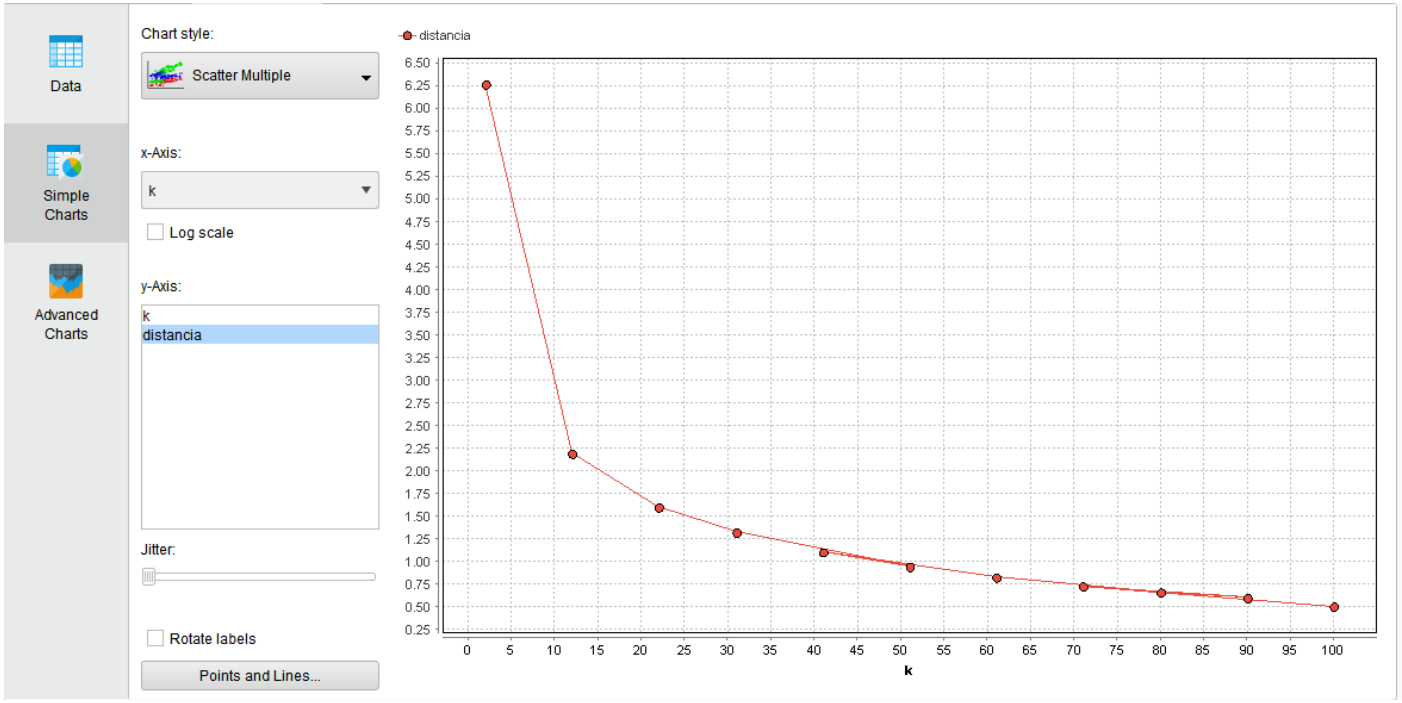
Através da ferramenta de clustering pretendeu-se fazer a segmentação dos dados relativos a clientes, nomeadamente:

- Género
- Idades
- Posição Financeira



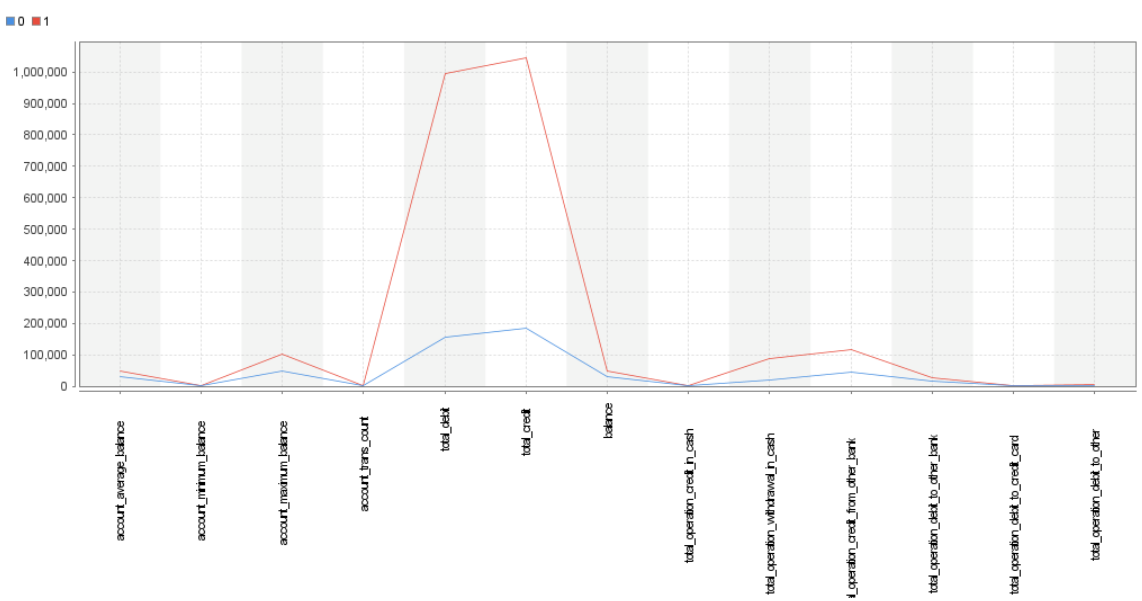
Análise Descritiva

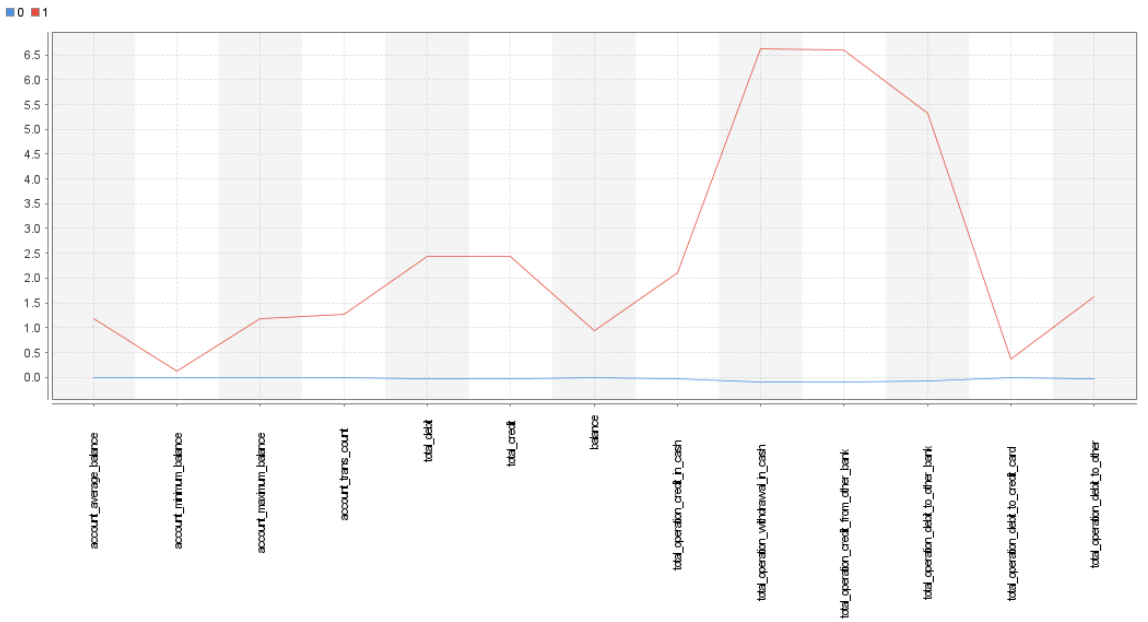
- Determinação do melhor k para o clustering, utilizando o "Elbow method"



Análise Descritiva

- Normalização dos atributos para atenuar escalas grandes ou "ouliers"
- Neste caso em concreto, sem normalização apenas poderiam ser evidenciados os atributos totais de movimentos a débito e totais de movimentos a crédito.
- Com a normalização tornou-se mais facil destacar os máximos, minimos, medias e desvios dos restantes atributos

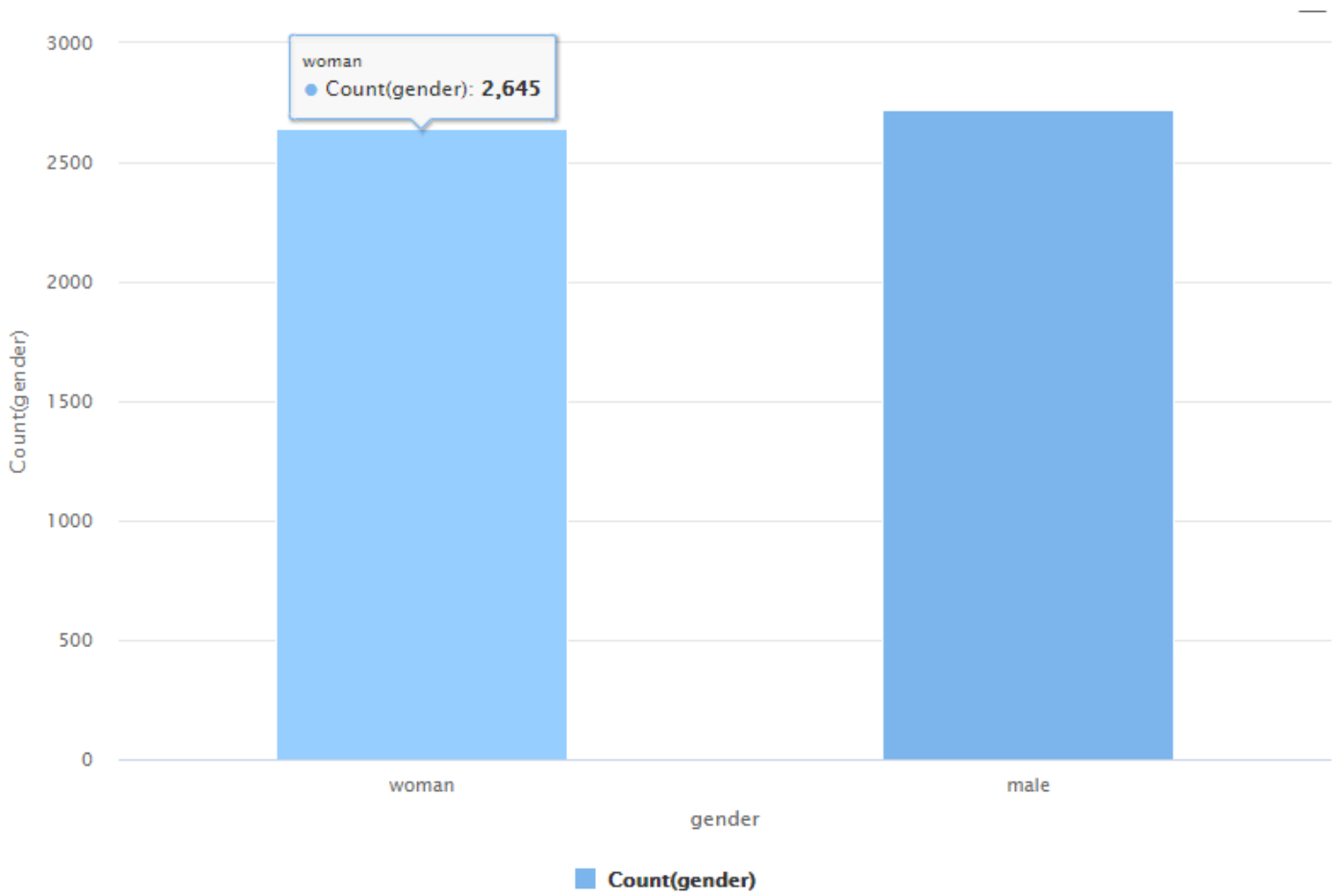




Análise Descritiva

- Relativamente ao género dos clientes estes são praticamente divididos igualmente entre homens e mulheres.

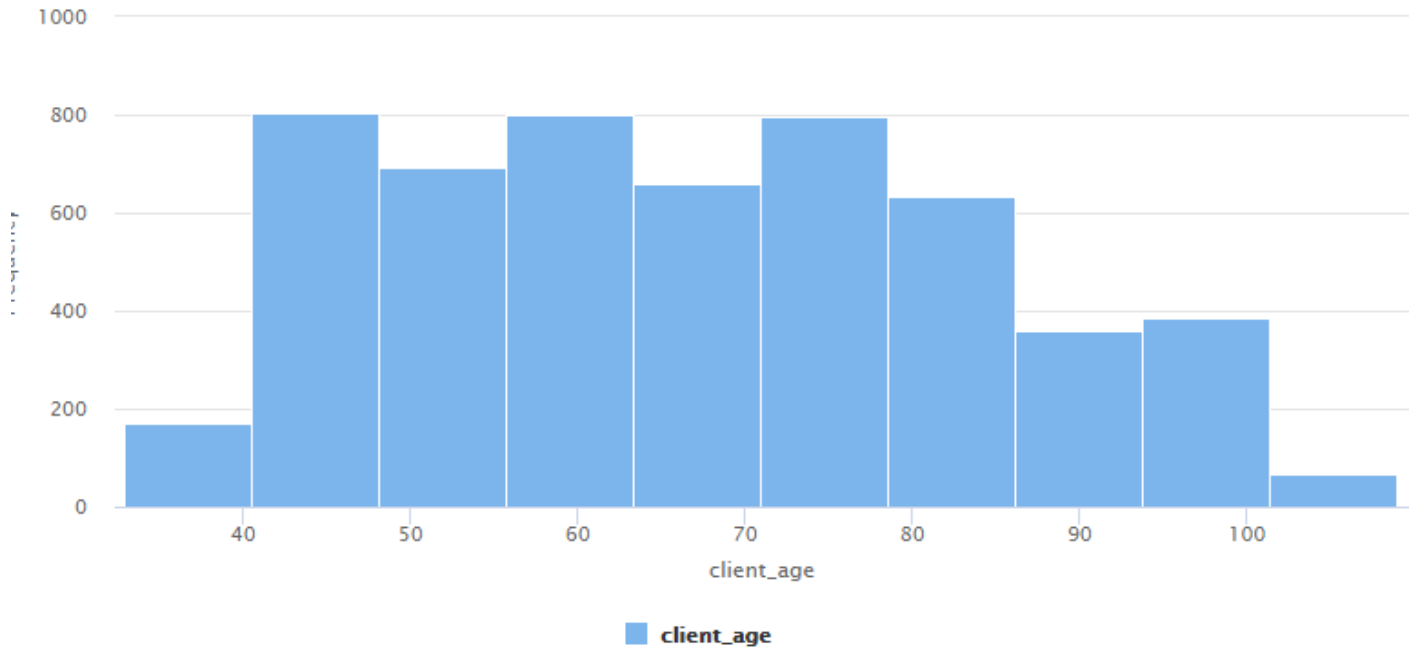
- O número de homens é 2724 e o número de mulheres é 2645.



Análise Descritiva

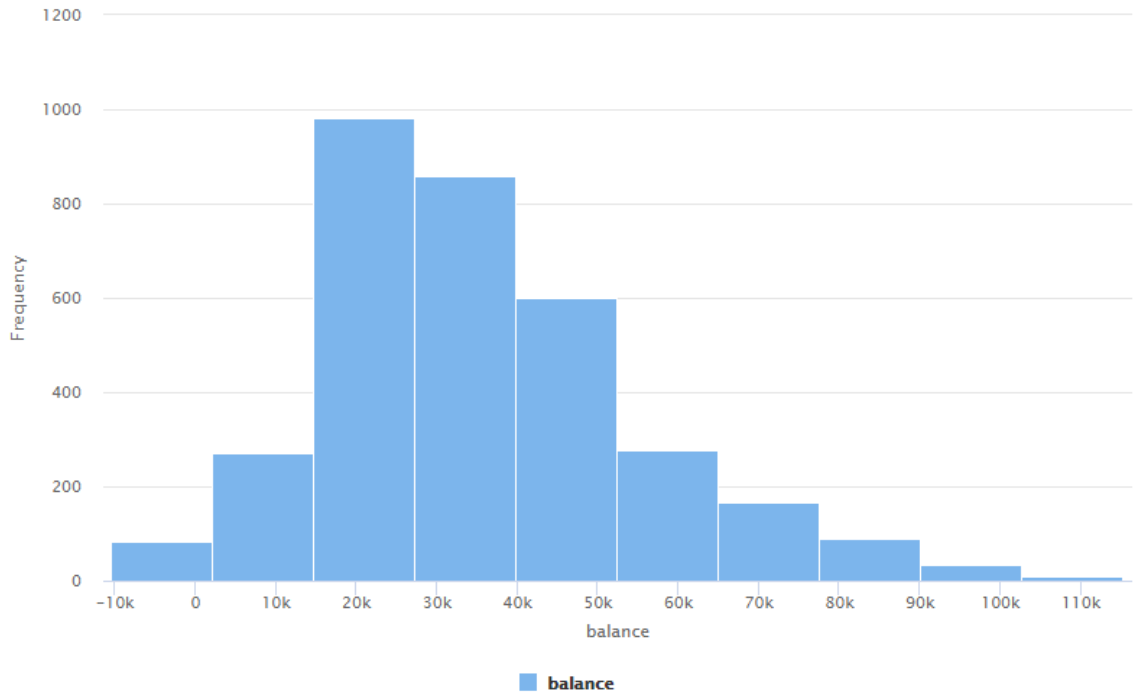
- A maioria da população em análise tem idade compreendida entre 40 a 100 anos.

- Sendo a idade média de: 66,77 anos

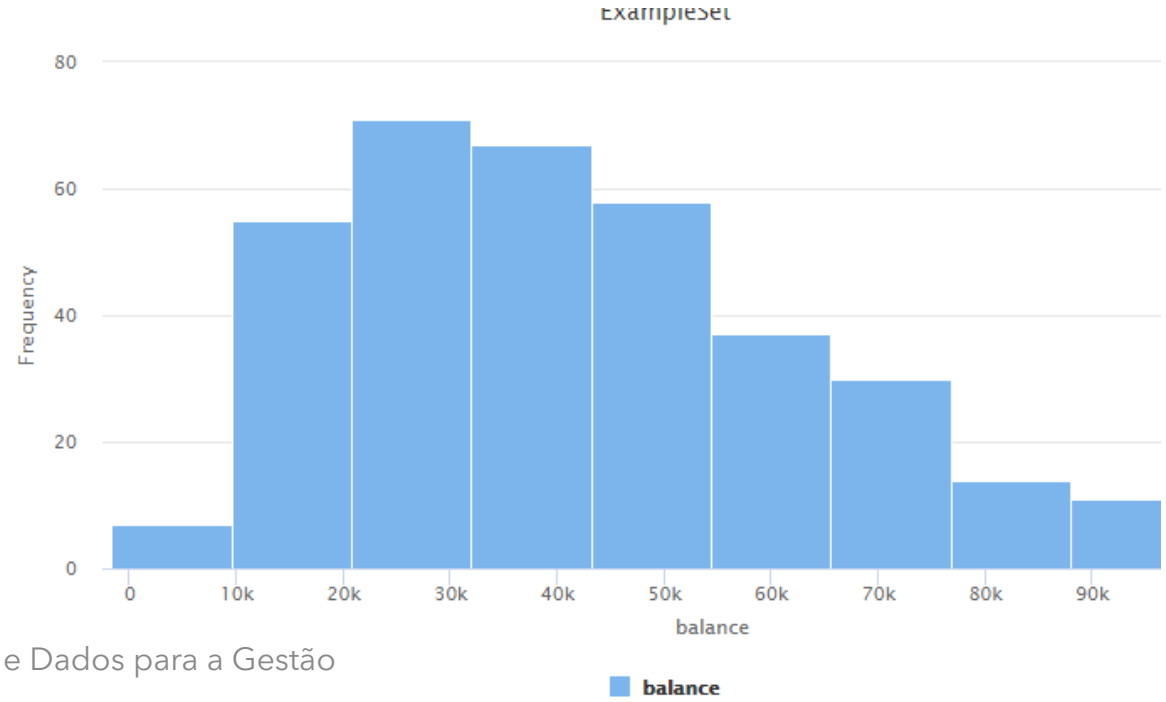


Análise Descritiva

- A maioria das contas bancárias de treino apresenta um saldo bancário entre os 20k e os 40k.
- O saldo médio é de 35,31k



- A maioria das contas bancárias de teste apresenta um saldo bancário entre os 30k e os 40k.
- O saldo médio é de 42,40k



Clustering

- Utilizando a técnica de clustering, e dividindo os dados em k=2 pode analisar-se algumas relações interessantes relativamente aos saldos médios, mínimos e máximos.
- Pode observar-se também que o salario médio não tem relação direta com o saldo médio.

Attribute	cluster_0	cluster_1
account_average_balance	28765.105	48996.021
account_minimum_balance	706.630	557.738
account_maximum_balance	49481.359	102977.329
account_trans_count	92.811	222.103
total_debit	154538.806	995533.329
total_credit	186425.107	1045202.107
balance	31886.301	49668.778
total_operation_credit_in_cash	1285.263	1134.361
total_operation_withdrawal_in_cash	20569.195	88226.240
total_operation_credit_from_other_bank	44184.664	117472.963
total_operation_debit_to_other_bank	17720.875	25649.783
total_operation_debit_to_credit_card	240.463	3049.307
total_operation_debit_to_other	1986.557	7057.815

Attribute	cluster_0	cluster_1
account_average_balance	48071.227	22400.557
average salary	9445.932	9442.822

Clustering

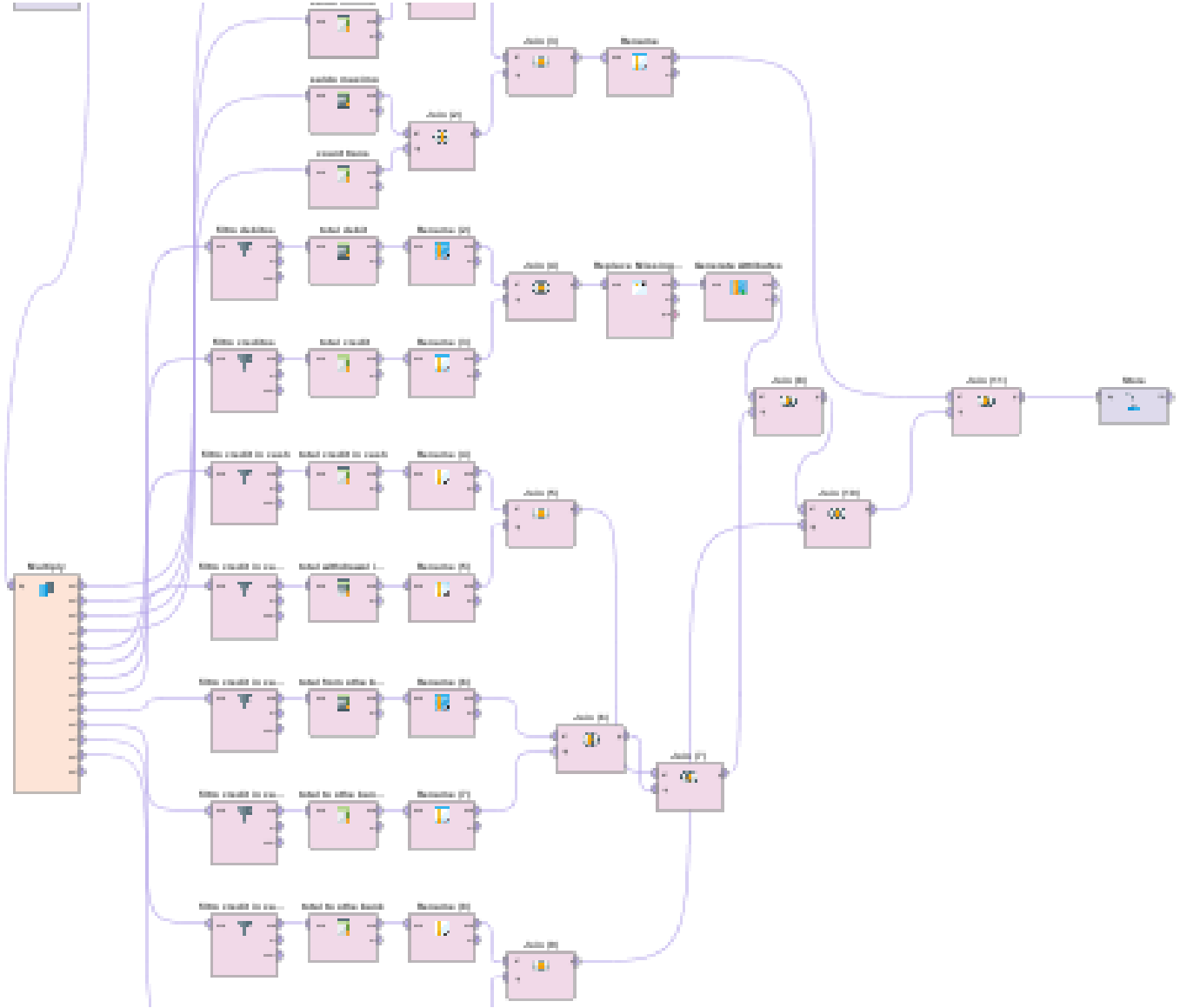
- Neste exemplo podem observar-se 2 clusters, relacionando os atributos saldo e o resultado do pagamento do empréstimo (dados de treino).

- Importância do atributo Saldo no pagamento dos empréstimos. **Normalmente, quem tem saldos maiores paga os empréstimos.**

4. ANÁLISE PREDITIVA

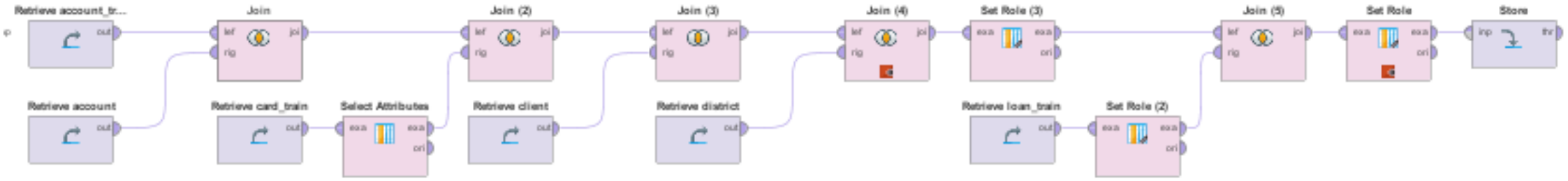
Análise Preditiva

Cálculo de novos atributos de contas bancárias (saldos, totais de movimentos) com operadores “aggregation”.



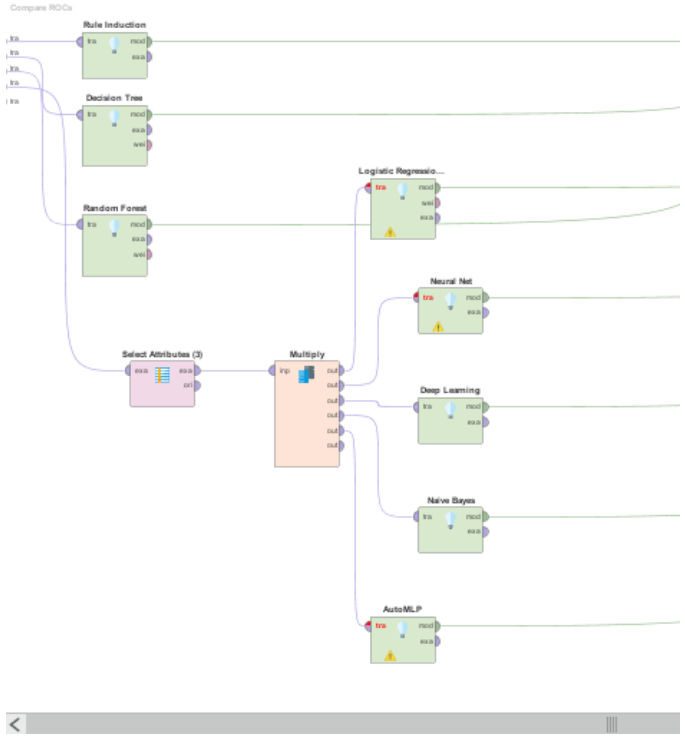
Análise Preditiva

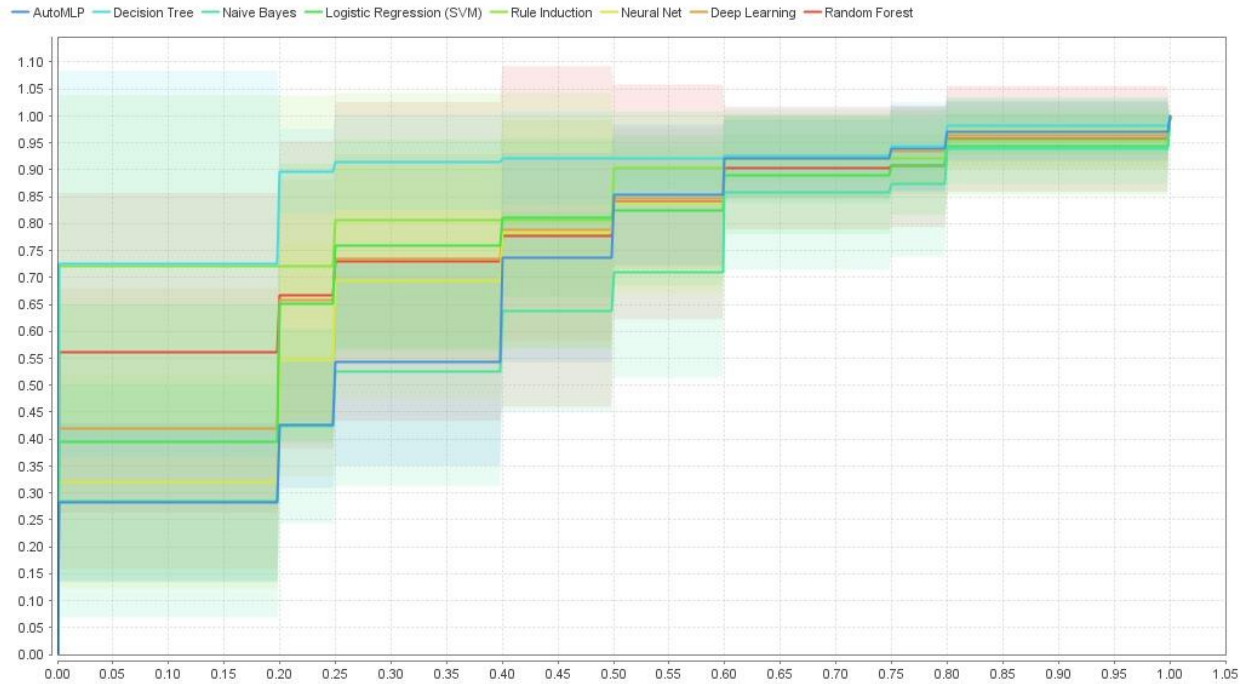
Agrupamento da Informação com operadores “Join”



Análise Preditiva

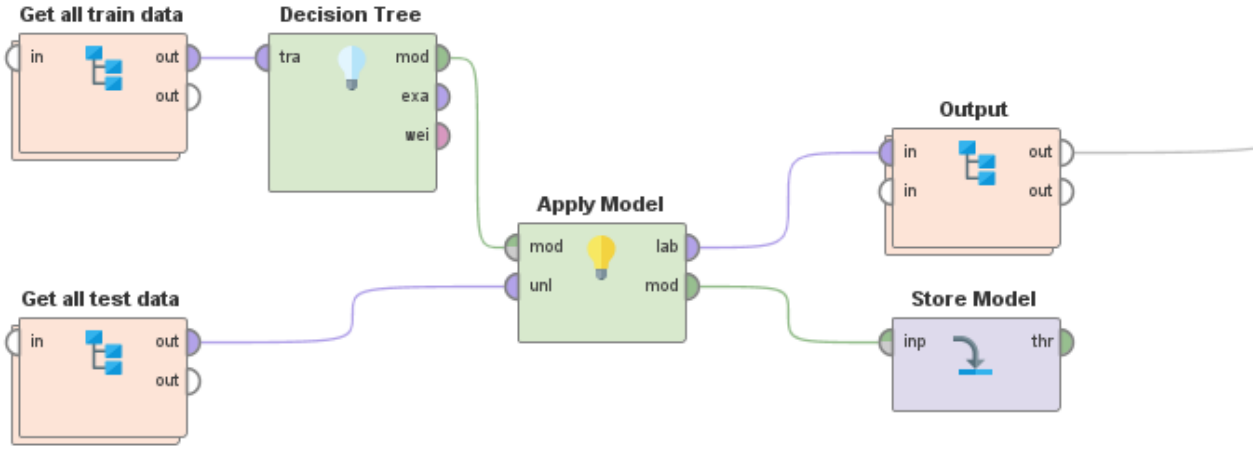
- A curva ROC (Receiver Operating Curve) representa a taxa de verdadeiros positivos (TVP) em função da taxa de falsos positivos (TFP).
- Comparação dos vários modelos de previsão, utilizando o melhor ROC como critério de seleção.
- Neste caso, a árvore de decisão (decision tree) apresentou os melhores resultados.

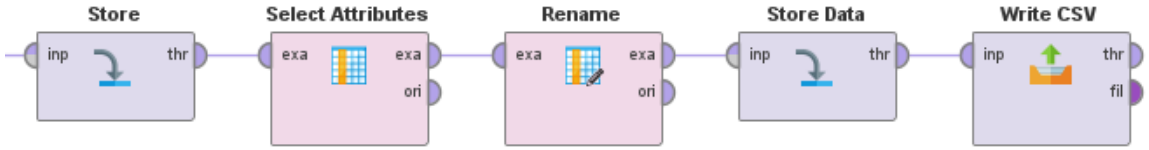


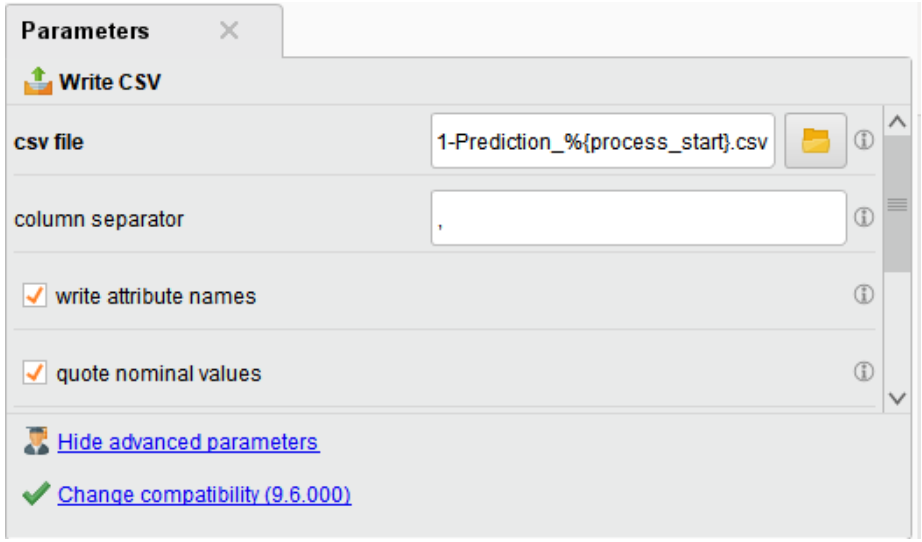


Análise Preditiva

- Criação do modelo de previsão (decision tree)
- Geração dos resultados finais utilizando parâmetros do sistema %{process_start} para a criação do CSV.







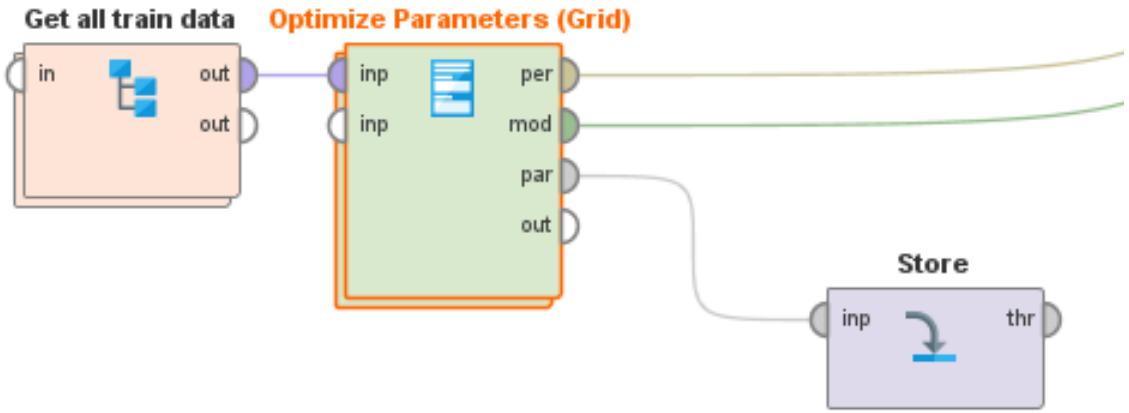
Análise Preditiva

- Otimização dos parâmetros da decision tree

ParameterSet

```
Parameter set:

Performance:
PerformanceVector [
-----accuracy: 96.92%
ConfusionMatrix:
True:    -1      1
-1:      7       0
1:       2       56
]
Decision Tree (2).criterion     = gain_ratio
Decision Tree (2).maximal_depth = 9
Decision Tree (2).minimal_leaf_size     = 1
Decision Tree (2).minimal_size_for_split        = 70
Decision Tree (2).confidence    = 0.10000008
```



Análise Preditiva

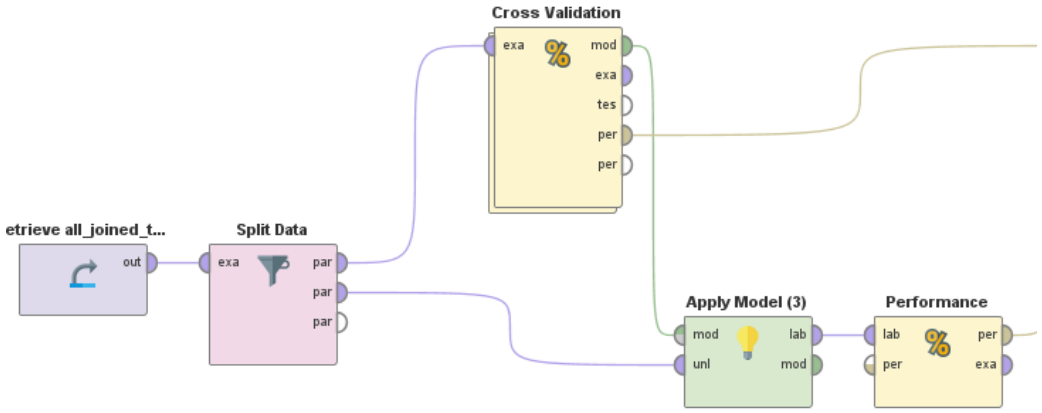
- Visualização dos detalhes da decision tree

Tree

```
account_minimum_balance > 54.050
|   account_average_balance > 19707.448
|   |   account_maximum_balance > 179279.250: -1 {-1=1, 1=0}
|   |   account_maximum_balance ≤ 179279.250
|   |   |   amount > 454692: -1 {-1=2, 1=2}
|   |   |   amount ≤ 454692
|   |   |   |   account_trans_count > 12.500
|   |   |   |   |   total_credit > 67937.450
|   |   |   |   |   |   balance > 16036.550
|   |   |   |   |   |   |   client_birthdate > Feb 1, 1938 11:30:00 PM GMT
|   |   |   |   |   |   |   |   account_maximum_balance > 158108.450: 1 {-1=1, 1=2}
|   |   |   |   |   |   |   |   account_maximum_balance ≤ 158108.450
|   |   |   |   |   |   |   |   |   client_birthdate > Jun 28, 1939 12:00:00 AM BST
|   |   |   |   |   |   |   |   |   |   account_trans_count > 13.500
|   |   |   |   |   |   |   |   |   |   |   total_credit > 77787.750: 1 {-1=7, 1=247}
|   |   |   |   |   |   |   |   |   |   |   total_credit ≤ 77787.750: -1 {-1=1, 1=1}
|   |   |   |   |   |   |   |   |   |   account_trans_count ≤ 13.500: 1 {-1=1, 1=3}
|   |   |   |   |   |   |   |   |   client_birthdate ≤ Jun 28, 1939 12:00:00 AM BST: 1 {-1=2, 1=5}
|   |   |   |   |   |   |   |   client_birthdate ≤ Feb 1, 1938 11:30:00 PM GMT: 1 {-1=2, 1=4}
|   |   |   |   |   |   balance ≤ 16036.550: 1 {-1=5, 1=10}
|   |   |   |   |   total_credit ≤ 67937.450: -1 {-1=3, 1=3}
|   |   |   |   account_trans_count ≤ 12.500: 1 {-1=4, 1=5}
|   account_average_balance ≤ 19707.448: -1 {-1=4, 1=0}
account_minimum_balance ≤ 54.050: -1 {-1=13, 1=0}
```

Análise Preditiva

- Medição da performance do modelo
- Cross Validation: 91,19% +-3,26%
- Performance Vector: 93,94%
- Split Data: 90% + 10%



accuracy: 93.94%

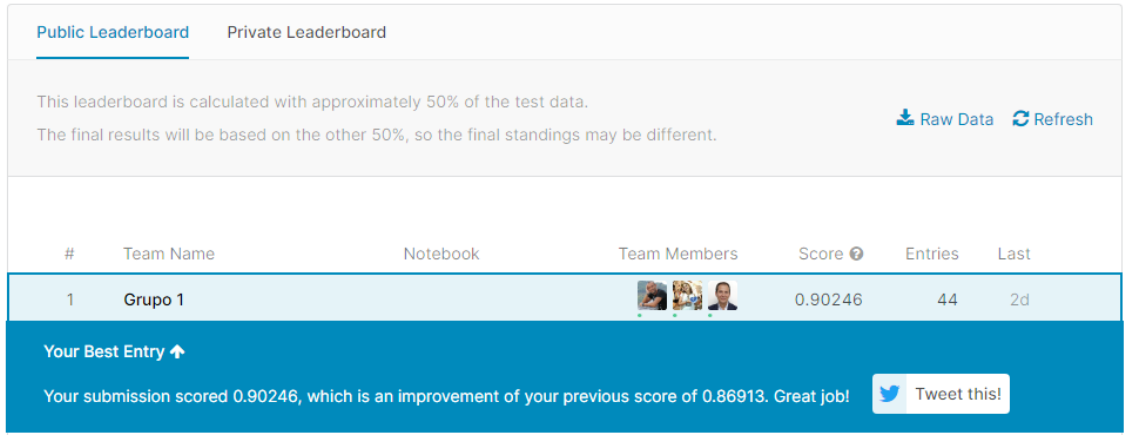
	true -1	true 1	class precision
pred. -1	3	0	100.00%
pred. 1	2	28	93.33%
class recall	60.00%	100.00%	

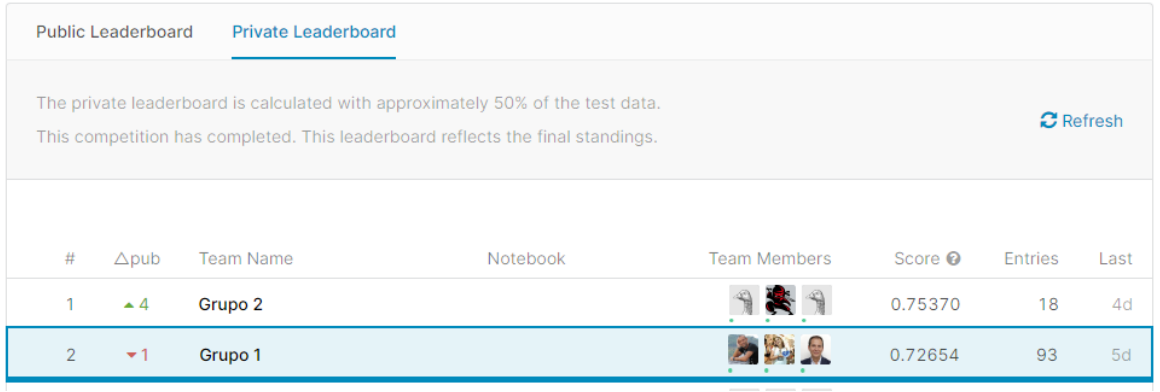
accuracy: 91.19% +/- 3.26% (micro average: 91.19%)

	true -1	true 1	class precision
pred. -1	21	6	77.78%
pred. 1	20	248	92.54%
class recall	51.22%	97.64%	

Análise Preditiva

- Pontuação na competição kaggle
- A discrepância de resultados justifica-se pelo overfit do modelo, perante dados não conhecidos





5. CONCLUSÕES, LIMITAÇÕES E TRABALHO FUTURO

CONCLUSÕES, LIMITAÇÕES E TRABALHO FUTURO

Os objetivos deste trabalho foram:

- Encontrar relações entre os atributos através da aplicação de técnicas de aprendizagem automática de dados;
- Criar modelos "inteligentes" capazes de auxiliar os bancos na tomada de decisão quanto à concessão de empréstimos aos seus clientes.

CONCLUSÕES, LIMITAÇÕES E TRABALHO FUTURO

A escolha dos atributos relevantes foi essencial para criar um modelo com a maior precisão possível.

A escolha errada de atributos ou em grande número, origina o "overfitting" provocando a degradação da performance do modelo e originando muitos falsos positivos e falsos negativos.

A comparação entre os diferentes algoritmos revelou que a árvore de decisão apresenta o melhor ROC.

Os dados de treino deveriam ser em maior número e cobrir mais anos de análise para melhorar a performance do modelo de previsão.

No futuro seria interessante estudar os dados relativos a mais anos de treino e comparar as respetivas previsões.

OBRIGADO !