ENG1034 - DATA MINING

RELATÓRIO DA AVALIAÇÃO 1

Carlos Mattoso

Matrícula: **1210553**

Rio de Janeiro, 6 de Maio de 2016

Introdução

Este trabalho apresenta uma análise e série de pré-processamentos da base de dados *buy,* com o objetivo de se desenvolver um modelo de aprendizado através do método *J48* capaz de classificar corretamente com um bom percentual de instâncias de tal base. Através dos pré-processamentos procurou-se elevar a taxa de acerto com relação a um modelo *J48* treinado sobre a base de treinamento original.

Primeiramente, descreve-se em detalhes a base de dados, seguida de uma análise exploratória da base de treinamento através da qual procurou-se fundamentar os pré-processamentos realizados. Finalmente, estes são exibidos na ordem que foram feitas com os resultados de modelos *J48* treinados sobre a respectiva base e comparados ao modelo controle. Para a realização deste trabalho utilizaram-se as ferramentas *Weka,* principalmente, e *R*, para funções auxiliares.

Descrição da Base de Dados

Descrição Qualitativa

A descrição da base de dados é feita em parte com base no arquivo *buy.pdf* disponibilizado no material da disciplina. Segundo este documento, a base indica no atributo *class* se o indivíduo em questão teria respondido ou não a uma promoção recente.

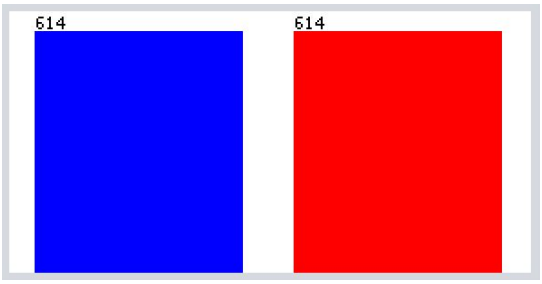
Nome do Atributo	Escala	Cardinalidade	Descrição
class	Nominal	Binária	1, caso tenha respondido 0, caso contrário
AGE	Razão	Discreta	Idade em anos
INCOME	Razão	Contínua	Renda anual
GENDER	Nominal	Binária	M, para masculino F, para feminino
MARRIED	Nominal	Binária	1, caso casado 0, caso contrário
FICO	Razão	Contínua	Nota de crédito americana do indivíduo
OWNHOME	Nominal	Binária	1, caso tenha casa própria 0, caso contrário
LOCATION	Nominal	Discreta	Local de residência, codificado de A a H
BUY6	Razão	Discreta	Número de compras nos últimos 6 meses
BUY12	Razão	Discreta	Número de compras nos últimos 12 meses
BUY18	Razão	Discreta	Número de compras nos últimos 18 meses
VALUE24	Razão	Contínua	Valor total de compras feitas nos últimos 24 meses
MOVED	Nominal	Binária	1, caso tenha se mudado nos últimos 6 meses 0, caso contrário

Descrição Quantitativa

Apresenta-se agora, para cada atributo, a distribuição dos valores da base de treinamento, incluindo-se as estatísticas de média, mediana, mínimo, máximo, número de *missing values* e frequência, segundo o que for aplicável. Além disso, exibem-se visualizações apropriadas dos atributos para facilitar sua compreensão. Aqui já foi feito um mínimo pré-processamento: foram alterados os tipos dos atributos *MARRIED, MOVED* e *OWNHOME* de *numeric* para *{0,1}* no arquivo de entrada, de modo que fiquem consistentes com a semântica de seus dados.

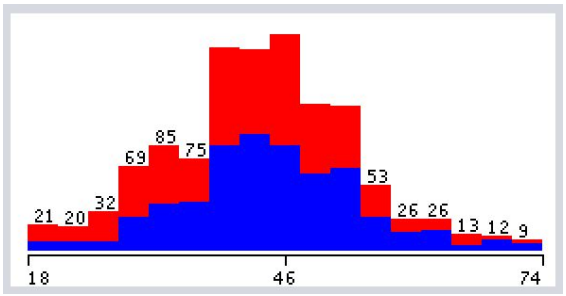
class

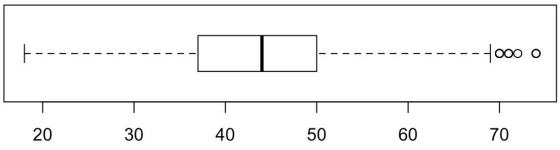
Valor	Frequência Absoluta	Frequência Relativa
0	614	50%
1	614	50%



AGE

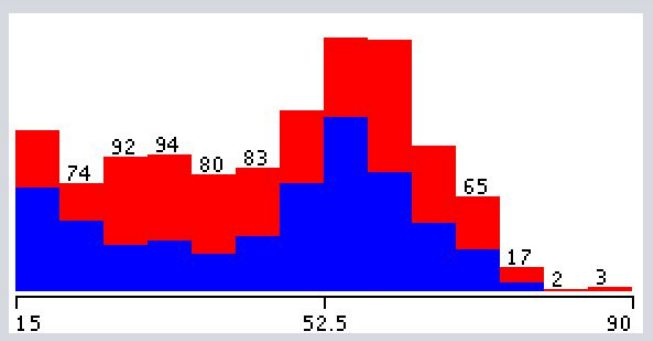
- Mínimo: 18.00
- Máximo: 74.00
- Mediana: 44.00
- Média: 43.472
- Desvio padrão: 10.065
- Missing: 44

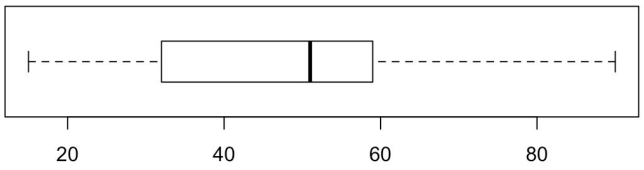




INCOME

- Mínimo: 15.00
- Máximo: 90.00
- Mediana: 51.00
- Média: 46.509
- Desvio padrão: 16.54
- Missing: 44

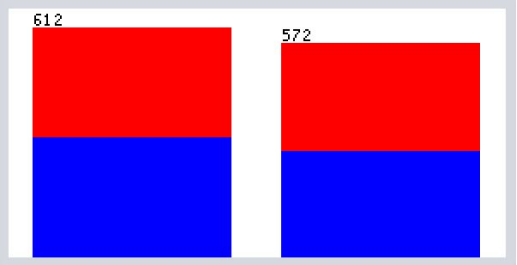




GENDER

Valor	Frequência Absoluta	Frequência Relativa
M	612	51.69%
F	572	48.31%

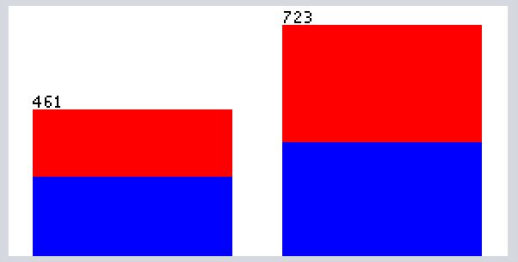
Missing: 44



MARRIED

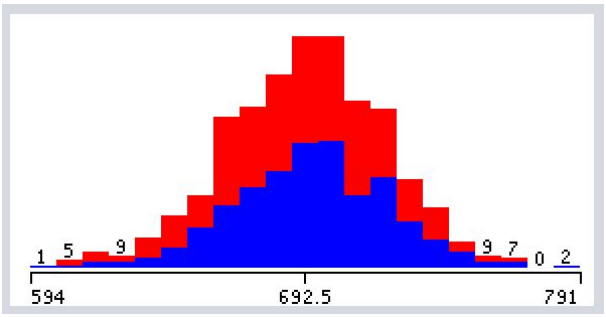
Valor	Frequência Absoluta	Frequência Relativa
0	461	38.93%
1	723	61.07%

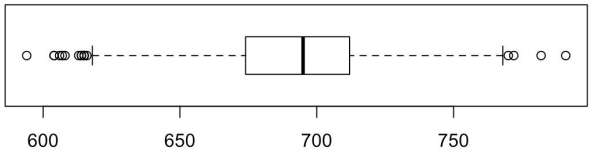
Missing: 44



FICO

- Mínimo: 594
- Máximo: 791
- Mediana: 695.0
- Média: 692.82
- Desvio padrão: 28.636
- Missing: 5





OWNHOME

Valor	Frequência Absoluta	Frequência Relativa
0	862	72.81%
1	322	27.19%

Missing: 44

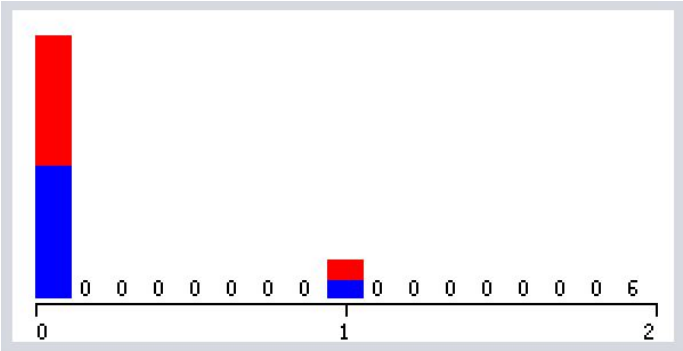
LOCATION

Valor	Frequência Absoluta	Frequência Relativa
A	21	1.71%
B	136	11.07%
C	68	5.54%
D	77	6.27%
E	206	16.78%
F	120	9.77%
G	178	14.50%
H	422	34.36%

Missing: 0

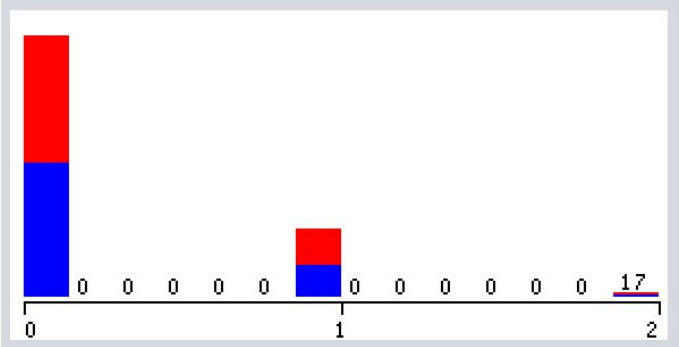
BUY6

- Mínimo: 0
- Máximo: 2
- Mediana: 0
- Média: 0.135
- Desvio padrão: 0.356
- Missing: 0



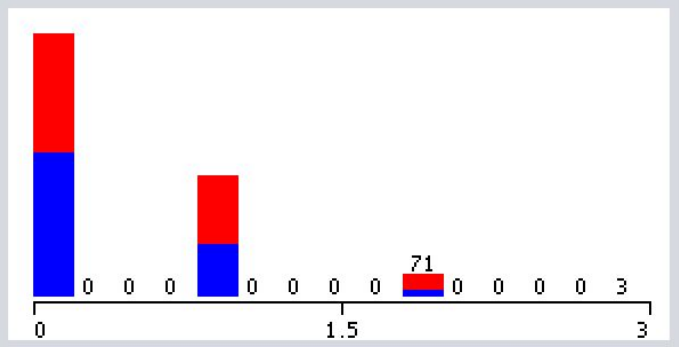
BUY12

- Mínimo: 0
- Máximo: 2
- Mediana: 0
- Média: 0.231
- Desvio padrão: 0.453
- Missing: 0



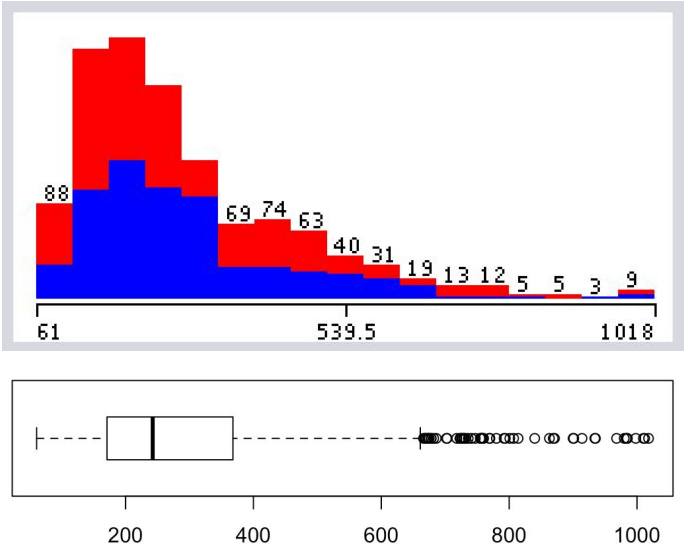
BUY18

- Mínimo: 0
- Máximo: 3
- Mediana: 0
- Média: 0.419
- Desvio padrão: 0.612
- Missing: 0



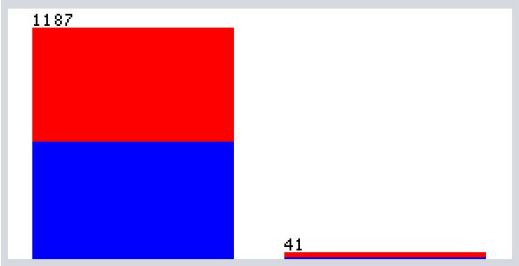
VALUE24

- Mínimo: 61
- Máximo: 1018
- Mediana: 242.5
- Média: 292.587
- Desvio padrão: 172.914
- Missing: 0



MOVED

Valor	Frequência Absoluta	Frequência Relativa
0	1187	96.66%
1	41	3.34%

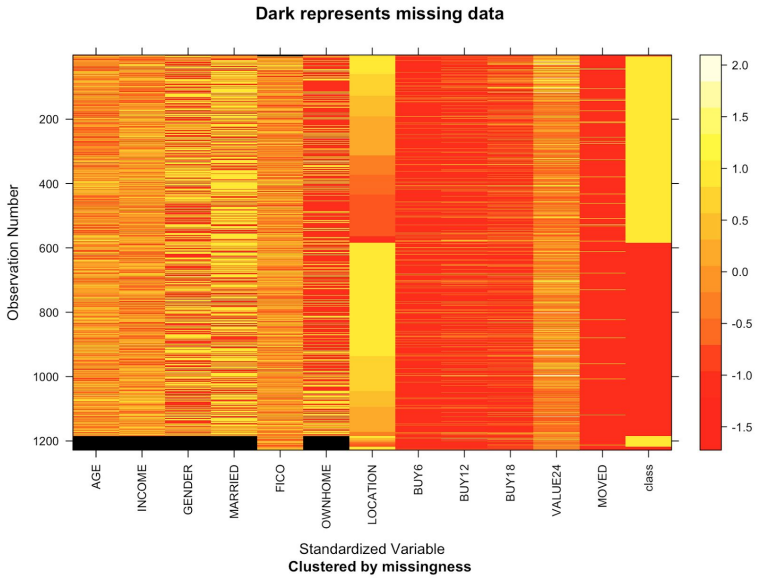


Missing: 0

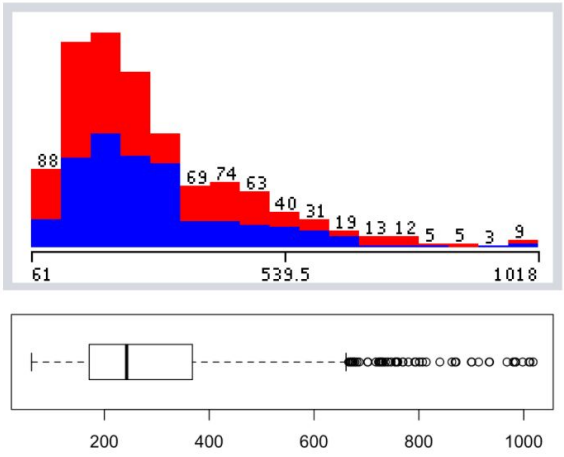
Análise Exploratória dos Dados

Podem ser extraídas observações interessantes com base nos gráficos apresentados acima. Tratemos de *missing values,* da forma de algumas distribuições e, por fim, de *outliers*.

Quanto a *missing values,* procurei explorar se estes se apresentam bem distribuídos ou se há algum padrão. Para visualizar isto, é útil o emprego de um mapa de *missing*, também chamado de *missing pattern plot*. Como pode-se ver ao lado, revela-se um claro padrão que pode ser atacado no desenvolvimento do modelo.



Este gráfico agrupa as unidades de observação que apresentam *missing values*. Note ainda que, um subconjunto dos registros tem praticamente todos os *missing values* existentes no banco de dados, com exceção dos de *FICO*. Sendo assim, pode-se tentar algumas estratégias para se lidar com isto: simplesmente deixá-los como estão, eliminar tais valores ou aplicar algum método de imputação múltipla.



Além disso, quanto a *outliers*, destaca-se o atributo *VALUE24*. Como pode-se ver em seu *boxplot*, este atributo apresenta uma quantidade significativa de valores extremos. Isto não necessariamente é algo prejudicial ao aprendizado, visto que tais *outliers* podem ser necessários para que o modelo capture uma região específica da distribuição. Afinal, é plausível que exista uma correlação (ou

vice-versa) entre pessoas cujo consumo nos últimos 24 meses tenha sido muito elevado e aquelas que responderam a promoção.

Ainda neste contexto, é válido aplicar um método de normalização, para colocar os dados em uma mesma escala. Creio que isto não terá muito efeito por estarmos usando um método determinístico linear de árvore, mas será investigado.

Finalmente, pode-se tentar simplificar um pouco o *dataset*. Como estamos lidando com um conjunto tão pequeno de instâncias, por volta de 1000 apenas, pode ser vantajoso discretizar alguns dos atributos, a fim de se generalizar um pouco mais as informações que destes o modelo é capaz de extrair. Esta e as demais estratégias serão exploradas na próximas seções.

Pré-processamento

Controle

Primeiramente, apresento o resultado de teste obtido pelo modelo treinado com a base sem pré-processamento, considerado como o controle para fins de comparação com os demais modelos produzidos. Além disso, para todos os casos o uso do algoritmo *J48* foi limitado aos parâmetros padrões.

```
  a   b   <-- classified as
 87  66 |   a = 0
 49 104 |   b = 1
```

```
Correctly Classified Instances         191               62.4183 %
Incorrectly Classified Instances       115               37.5817 %
```

Análise de *Missing Values*

Primeiramente, tentemos simplesmente remover todos as instâncias com algum *missing value*. Isto foi feito utilizando-se a função *complete.cases* de *R* para identificar todas as linhas com *missing values*. Então, o novo arquivo resultante foi carregado no *Weka* para avaliação:

```
  a   b   <-- classified as
 89  64 |   a = 0
 49 104 |   b = 1
```

```
Correctly Classified Instances         193               63.0719 %
Incorrectly Classified Instances       113               36.9281 %
```

Obtivemos uma pequena melhoria com relação ao modelo original. Agora vamos tentar dois métodos para imputação de *missing values*. Primeiramente, utiliza-se o método do *Weka, ReplaceMissingValues*, que realizada as substituições através de modas e médias:

```
  a   b   <-- classified as
 81  72 |   a = 0
 50 103 |   b = 1
```

```
Correctly Classified Instances         184               60.1307 %
Incorrectly Classified Instances       122               39.8693 %
```

Este método se saiu pior do que o controle. Outra tentativa foi usar o método *Multivariate Imputation by Chained Equations* (MICE) em *R,* disponibilizado pela biblioteca *mice*. Além disso, antes de realizar a imputação, segmentei os dados de acordo com a classe, para forçar uma imputação relativa a classe. Vejamos como a nova base se saiu:

```
 a   b   <-- classified as
81  72 |   a = 0
41 112 |   b = 1

Correctly Classified Instances         193               63.0719 %
Incorrectly Classified Instances       113               36.9281 %
```

Este modelo teve desempenho idêntico ao derivado da base sem as instâncias que tinham *missing value*. Embora suas demais estatísticas sejam levemente superiores (*ROC Area*, *RMSE*, etc), como exibido abaixo, opta-se por continuar com o segundo modelo. Além disso, isto revela que tais instâncias aparentemente não contribuem significativamente para enriquecer o modelo, visto que sua complementação com o restante do modelo não consegue superar o caso em que usamos apenas o restante.

```
Kappa statistic                          0.2614
Mean absolute error                      0.3829
Root mean squared error                  0.5429
Relative absolute error                 76.5767 %
Root relative squared error            108.5759 %
Total Number of Instances              306

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.529    0.268    0.664      0.529    0.589      0.267    0.669     0.608     0
                 0.732    0.471    0.609      0.732    0.665      0.267    0.669     0.668     1
Weighted Avg.    0.631    0.369    0.636      0.631    0.627      0.267    0.669     0.638
```

Imputação com MICE

```
Kappa statistic                          0.2614
Mean absolute error                      0.3838
Root mean squared error                  0.5585
Relative absolute error                 76.753  %
Root relative squared error            111.6801 %
Total Number of Instances              306

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.582    0.320    0.645      0.582    0.612      0.263    0.665     0.587     0
                 0.680    0.418    0.619      0.680    0.648      0.263    0.665     0.672     1
Weighted Avg.    0.631    0.369    0.632      0.631    0.630      0.263    0.665     0.630
```

Remoção de instâncias com *missing*

Análise de *Outliers*

O filtro *InterquartileRange* no *Weka* não apontou *outliers* na base sem *missing*. Sendo assim, visando ainda verificar o impacto dos *outliers* no modelo, alterei mais uma vez a base de dados para removê-los, segundo uma relaxação do critério do filtro. Adotei um critério que verifica apenas a condição inferior: **`Q3 + OF*IQR < x <= Q3 + EVF*IQR`**. Contudo, como pode-se observar abaixo isto piorou bastante os resultados, sendo esta mudança então descartada.

```
 a  b   <-- classified as
85 68 |  a = 0
58 95 |  b = 1
```

```
Correctly Classified Instances         180               58.8235 %
Incorrectly Classified Instances       126               41.1765 %
```

Análise de Normalização

É também válido, como abordado na análise exploratória, avaliar uma possível normalização do atributo *VALUE24*. Isto porque ele apresenta o mais alto desvio padrão de todos os demais atributos e uma distribuição fortemente *right skewed*. Este atributo foi então normalizado, executando-se um filtro *Standartize* em modo *Batch,* para que a normalização fosse comum ao bancos de treinamento e de teste ao mesmo tempo. Um inconveniente deste filtro é ele ser aplicado a todos os atributos numéricos.

```
  a   b   <-- classified as
 89  64 |   a = 0
 49 104 |   b = 1
```

```
Correctly Classified Instances         193               63.0719 %
Incorrectly Classified Instances       113               36.9281 %
```

Como havia suposto a princípio, isto não trouxe melhoria alguma. De fato, o algoritmo que estamos usando é linear e determinístico, isto é, dada uma mesma entrada (ou uma transformação desta), espera-se a mesma saída.

Análise de Discretização

Finalmente verifica-se a transformação de dados numéricos em categóricos para averiguar se este processo contribui para melhorias nos resultados. Primeiramente, retorna-se a base sem *missing* sem *standartization*, já que este último processo em nada havia aprimorado os resultados. Dos atributos numéricos que temos passíveis de discretização, os principais são *AGE, INCOME, FICO* e *VALUE24*. Para produzir bons resultados, foram necessários alguns experimentos. Apresento abaixo alguns resultados:

```
  a   b   <-- classified as
 92  61 |   a = 0
 44 109 |   b = 1
```

```
Correctly Classified Instances         201               65.6863 %
Incorrectly Classified Instances       105               34.3137 %
```

AGE, INCOME, FICO: 5 bins; **VALUE24**: 10 bins

```
  a   b   <-- classified as
 71  82 |   a = 0
 24 129 |   b = 1
```

```
Correctly Classified Instances         200               65.3595 %
Incorrectly Classified Instances       106               34.6405 %
```

AGE, INCOME: 5 bins; **FICO, VALUE24**: 6 bins

```
 a   b   <-- classified as
91  62 |   a = 0
39 114 |   b = 1

Correctly Classified Instances         205               66.9935 %
Incorrectly Classified Instances       101               33.0065 %
```

AGE, INCOME, FICO, VALUE24: 10 bins

Ao simplificar a base de dados através da discretização, atingiu-se quase 67%, de taxa de acerto, uma melhoria de 5 pontos percentuais frente ao resultado original.

Seleção de variáveis

Utilizando-se a base de dados pré-processada em que foram eliminadas instâncias com *missing values* e discretizadas as variáveis *AGE, INCOME, FICO* e *VALUE24*, apresenta-se o resultado do processo de seleção de variáveis.

Rank do Atributo	Métodos				
	1R	**Gain**	**InfoGain**	**Relief**	**Correlação**
1	LOCATION	LOCATION	LOCATION	LOCATION	LOCATION
2	INCOME	BUY18	INCOME	INCOME	BUY18
3	BUY18	MOVED	VALUE24	VALUE24	INCOME
4	VALUE24	INCOME	AGE	BUY18	MOVED
5	AGE	VALUE24	BUY18	GENDER	VALUE24
6	BUY12	AGE	FICO	OWNHOME	OWNHOME
7	MARRIED	FICO	MOVED	MOVED	MARRIED
8	MOVED	OWNHOME	OWNHOME	BUY12	BUY12
9	BUY6	MARRIED	MARRIED	MARRIED	AGE
10	GENDER	GENDER	GENDER	BUY6	GENDER
11	FICO	BUY12	BUY12	AGE	BUY6
12	OWNHOME	BUY6	BUY6	FICO	FICO

É nítido que há um consenso que os atributos *LOCATION, INCOME, BUY18* e *VALUE24* são de extrema relevância. Por outro lado, *GENDER, BUY6* e *FICO* são os que mais figuram no final do *rank*. Avaliemos agora a contribuição individual de cada um e conjunta, para chegarmos ao modelo final.

```
 a   b   <-- classified as
91  62 |   a = 0
39 114 |   b = 1

Correctly Classified Instances         205               66.9935 %
Incorrectly Classified Instances       101               33.0065 %
```

Sem *GENDER*

```
 a   b   <-- classified as
91  62 |   a = 0
39 114 |   b = 1

Correctly Classified Instances         205               66.9935 %
Incorrectly Classified Instances       101               33.0065 %
```

Sem *BUY6*

```
 a   b   <-- classified as
91  62 |   a = 0
39 114 |   b = 1

Correctly Classified Instances         205               66.9935 %
Incorrectly Classified Instances       101               33.0065 %
```

Sem *FICO*

```
 a   b   <-- classified as
91  62 |   a = 0
39 114 |   b = 1

Correctly Classified Instances         205               66.9935 %
Incorrectly Classified Instances       101               33.0065 %
```

Sem *GENDER, BUY6* e *FICO*

É nítido que a eliminação destes atributos em nada afeta os resultados. Portanto, podemos eliminá-los sem perda de poder preditivo.

Por completude, demonstra-se a relevância dos 3 principais atributos (*LOCATION, INCOME* e *BUY18)*. Seguem os resultados para um modelo treinado sobre uma base sem tais atributos:

```
 a  b   <-- classified as
85 68 |  a = 0
60 93 |  b = 1

Correctly Classified Instances         178               58.1699 %
Incorrectly Classified Instances       128               41.8301 %
```

A taxa de acerto que estava em aproximadamente 67% nos modelos anteriores, caiu 9 pontos percentuais! Uma queda bastante expressiva. Isto reforça os resultados dos algoritmos de seleção de variáveis e mostra a importância de tais atributos para o aprendizado do modelo.

Conclusão

Em conclusão, fui capaz de obter um modelo com taxa de acerto de aproximadamente 67%, produzido através da eliminação de *missing values*, discretização de 4 variáveis numéricas contínuas em 10 *bins* cada (*AGE, INCOME, FICO* e *VALUE24*) e com 3 atributos eliminados *(GENDER, BUY6* e *FICO)*. Embora tenha reduzido a dimensionalidade do *dataset* e aumentado em 5 pontos percentuais a taxa de acerto, me parece que o modelo chegou em um ponto em que aprimoramentos ficaram difíceis e em que seria melhor recomeçar e tentar um caminho diferente.

Além disso, seria interessante também estudar o algoritmo *J48* a fundo para experimentação com seus parâmetros a fim de aprimorar os modelos produzidos, ao invés de usar apenas seu modo padrão. Aplicar outros métodos que possam ser mais propícios a este *dataset* e compará-los ao *J48* também seria interessante.