ENG1034 - DATA MINING

RELATÓRIO DA AVALIAÇÃO 2

Carlos Mattoso

Matrícula: **1210553**

Rio de Janeiro, 24 de Junho de 2016

Introdução

Este trabalho apresenta uma análise e série de pré-processamentos da base de dados *buy,* com o objetivo de se desenvolver e comparar diferentes modelos de aprendizado através dos métodos *Naive*, *KNN (IBK)*, *PART* e *J48,* visando 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 uma seleção de modelos treinados 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 feitos com os resultados dos modelos 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.

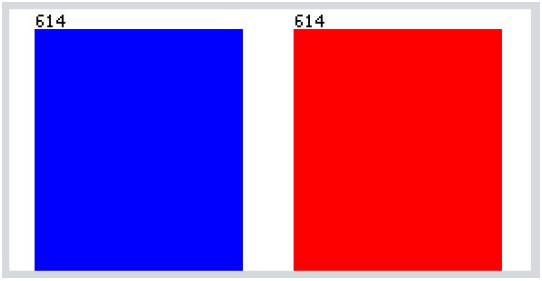
Nome do Atributo	Escala	Cardinalidade	Descrição
class	Nominal	Binária	1, caso tenha respondido 0, caso contrário
AGE	Razão	Discreta	Idade em anos
INCOME	Razão	Contínua	Renda anual
GENDER	Nominal	Binária	M, para masculino F, para feminino
MARRIED	Nominal	Binária	1, caso casado 0, caso contrário
FICO	Razão	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. As bases de treinamento e teste resultantes desta modificação são considerada as bases de dados originais, e refere-se a elas desta forma ao longo deste trabalho.

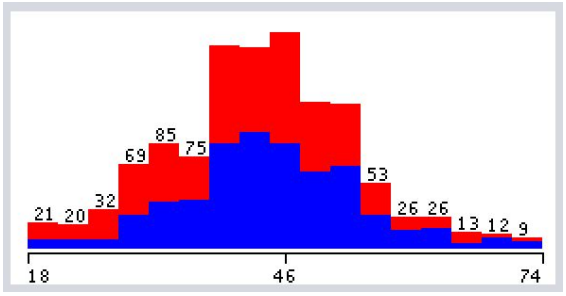
class

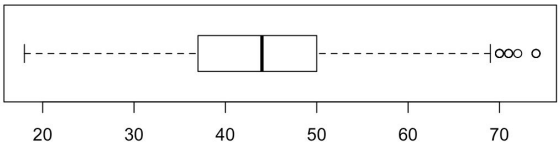
Valor	Frequência Absoluta	Frequência Relativa
0	614	50%
1	614	50%



AGE

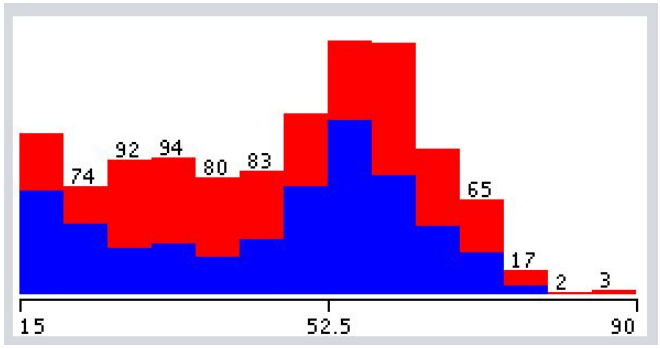
- Mínimo: 18.00
- Máximo: 74.00
- Mediana: 44.00
- Média: 43.472
- Desvio padrão: 10.065
- Missing: 44

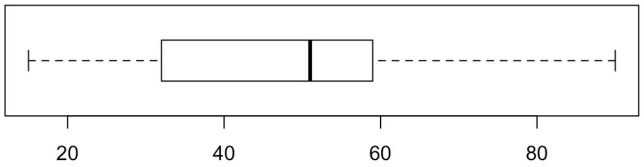




INCOME

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

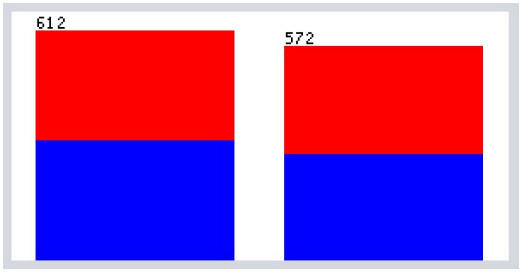




GENDER

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

Missing: 44



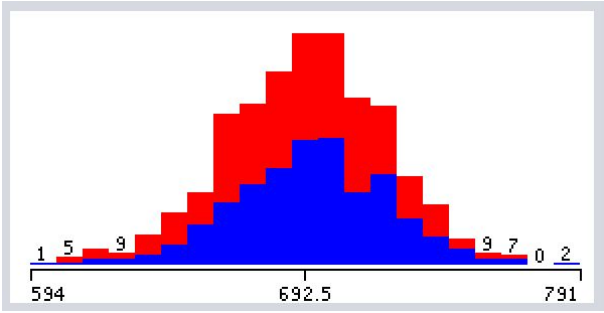
MARRIED

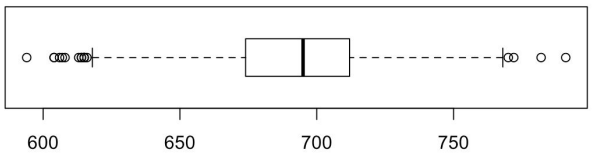
Valor	Frequência Absoluta	Frequência Relativa
0	461	38.93%
1	723	61.07%

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

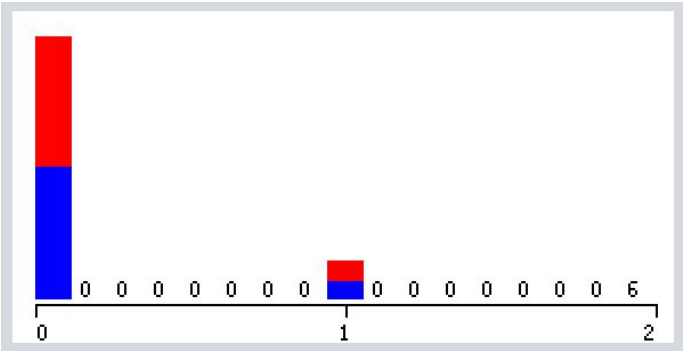
Valor	Frequência Absoluta	Frequência Relativa
A	21	1.71%
B	136	11.07%
C	68	5.54%
D	77	6.27%
E	206	16.78%

Missing: 0

F	120	9.77%
G	178	14.50%
H	422	34.36%

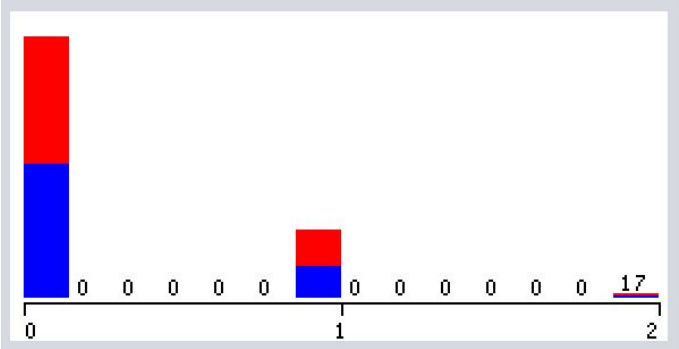
BUY6

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



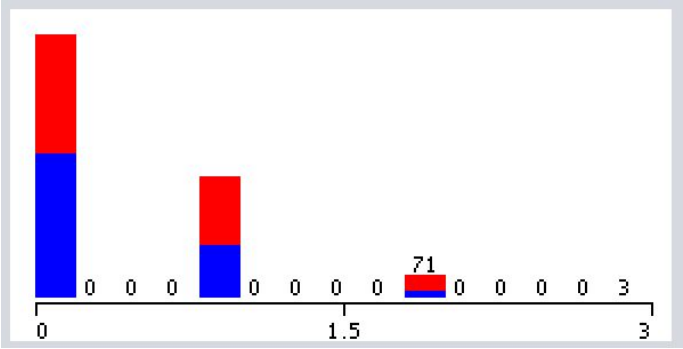
BUY12

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



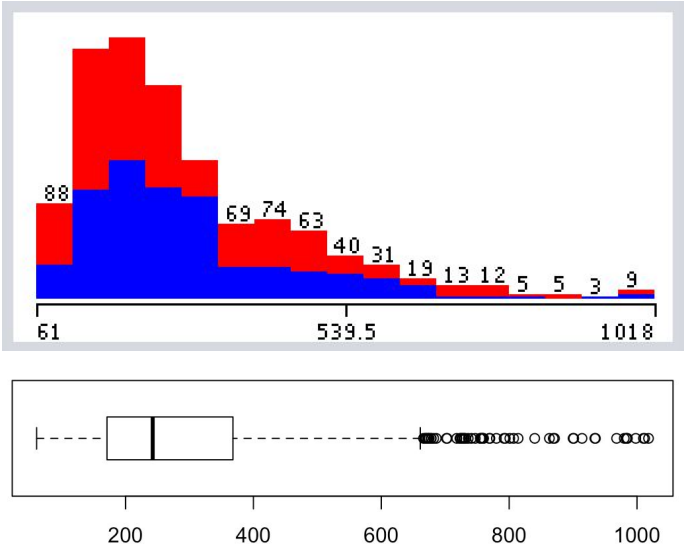
BUY18

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



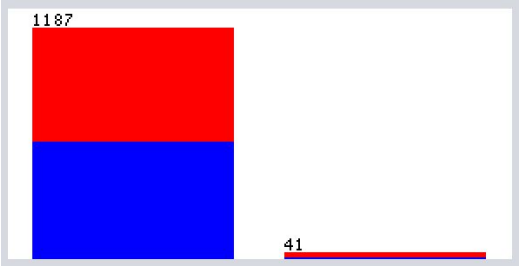
VALUE24

- 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%



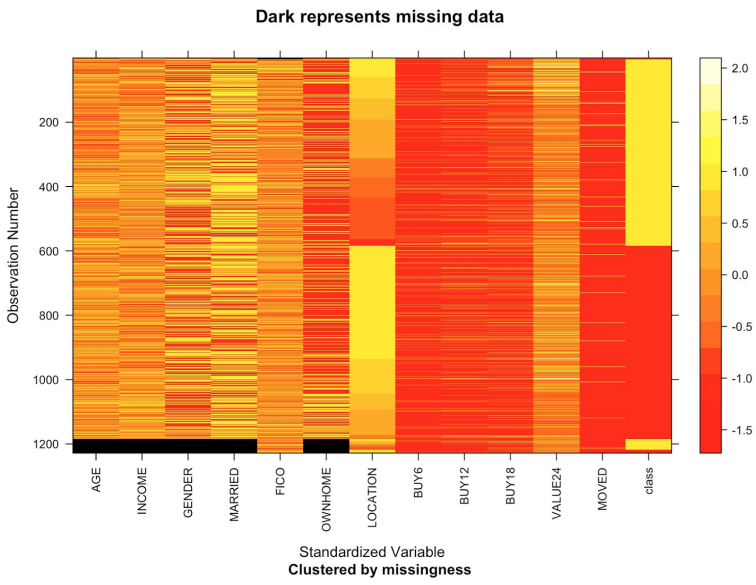
Missing: 0

Análise Exploratória dos Dados

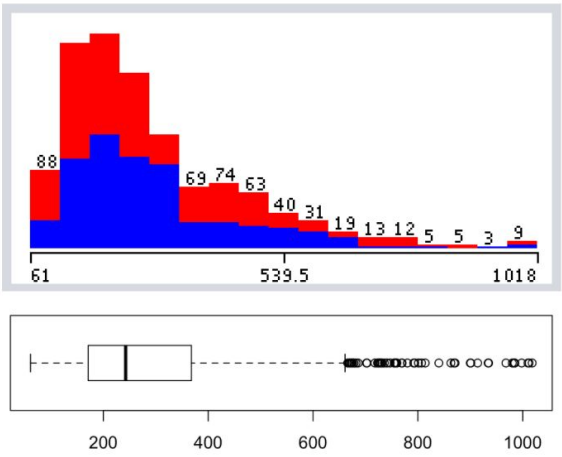
Podem ser extraídas observações interessantes com base nos gráficos apresentados acima. Tratemos de *missing values,* da forma de algumas distribuições e, por fim, de *outliers*.

Note que as variáveis *BUY6, BUY12* e *BUY18*, embora semanticamente indiquem intervalos como apontado no arquivo *buy.pdf* disponibilizado no material da disciplina, tem propriedades nítidas de variável categórica, devido a apresentarem um conjunto limitado de valores observados. Em razão disto, serão analisados os resultados alterando-se tais variáveis para categóricas.

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 os resultados de avaliação da base de testes obtidos pelo modelos dos diferentes métodos treinados com a base de treinamento original, sem pré-processamento. Estes resultados são considerados o controle para fins de comparação com os demais modelos produzidos. Para o *KNN*, testaram-se diferentes valores de *K* a fim de se selecionar um conjunto de valores de *K* para serem usados no restante do trabalho.

Algumas variações são experimentadas quanto aos parâmetros dos algoritmos, sendo mantida posteriormente a configuração que produziu os melhores resultados.

Naive

Correctly Classified Instances	198	64.7059 %
Incorrectly Classified Instances	108	35.2941 %

KNN (IBk)

Primeiro, exibem-se os resultados abaixo para diferentes valores de *K* com as configurações padrões. Posteriormente, variam-se os pesos dos votos dos vizinhos, utilizando-se o inverso da distância, de modo que quão mais próximo um vizinho mais significativo é seu voto na eleição de classe.

Ponderação Uniforme

k = 1

Correctly Classified Instances	192	62.7451 %
Incorrectly Classified Instances	114	37.2549 %

k = 3

Correctly Classified Instances	188	61.4379 %
Incorrectly Classified Instances	118	38.5621 %

k = 5

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

k = 7

Correctly Classified Instances	173	56.5359 %
Incorrectly Classified Instances	133	43.4641 %

Para *K* maiores constata-se que os resultados apenas pioram.

Ponderação por Inverso da Distância

k = 1

Correctly Classified Instances	192	62.7451 %
Incorrectly Classified Instances	114	37.2549 %

k = 3

Correctly Classified Instances	186	60.7843 %
Incorrectly Classified Instances	120	39.2157 %

k = 5

Correctly Classified Instances	183	59.8039 %
Incorrectly Classified Instances	123	40.1961 %

Os resultados pioraram, indicando *overfitting* a base de treinamento. Sendo assim, para análises posteriores será empregado o KNN com seus atributos padrões para *K = {1, 3, 5}.*

PART

Testam-se diferentes valores de *pruning* a fim de avaliar o que produz os melhores resultados.

C = 0.10

Correctly Classified Instances	203	66.3399 %
Incorrectly Classified Instances	103	33.6601 %

C = 0.25

Correctly Classified Instances	190	62.0915 %
Incorrectly Classified Instances	116	37.9085 %

C = 0.50

Correctly Classified Instances	190	62.0915 %
Incorrectly Classified Instances	116	37.9085 %

C = 0.75

Correctly Classified Instances	202	66.0131 %
Incorrectly Classified Instances	104	33.9869 %

C = 1.00

Correctly Classified Instances	202	66.0131 %
Incorrectly Classified Instances	104	33.9869 %

Os melhores resultado obtiveram-se para *C = 0.10,* que acarreta uma estratégia de *pruning* mais agressiva, e C = {0.75, 1.00}. Em análises posteriores este valor serão usados *0.10* e *0.75*.

J48

Aqui também testaram-se diferentes valores de *pruning*.

C = 0.10

```
Correctly Classified Instances         196               64.0523 %
Incorrectly Classified Instances       110               35.9477 %
```

C = 0.25

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

C = 0.50

```
Correctly Classified Instances         197               64.3791 %
Incorrectly Classified Instances       109               35.6209 %
```

C = 0.75

```
Correctly Classified Instances         194               63.3987 %
Incorrectly Classified Instances       112               36.6013 %
```

C = 1.00

```
Correctly Classified Instances         194               63.3987 %
Incorrectly Classified Instances       112               36.6013 %
```

Os melhores resultados obtiveram-se para *C* de 0.10, uma estratégia mais agressiva de *pruning* que a padrão, e de 0.50, uma estratégia menos agressiva. Ambos os valores serão usados em avaliações posteriores.

Ajuste das Variáveis *BUY*

Primeiramente, avaliemos os resultados obtidos alterando o tipo das variáveis *BUY6, BUY12* e *BUY18* para categóricas.

Naive

Correctly Classified Instances	195	63.7255 %
Incorrectly Classified Instances	111	36.2745 %

KNN (IBk)

k = 1

Correctly Classified Instances	190	62.0915 %
Incorrectly Classified Instances	116	37.9085 %

k = 3

Correctly Classified Instances	190	62.0915 %
Incorrectly Classified Instances	116	37.9085 %

k = 5

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

PART

C = 0.10

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

C = 0.75

Correctly Classified Instances	190	62.0915 %
Incorrectly Classified Instances	116	37.9085 %

J48

C = 0.10

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

C = 0.50

Correctly Classified Instances	195	63.7255 %
Incorrectly Classified Instances	111	36.2745 %

Conclusão

Os resultados obtidos foram piores do que os obtidos com a base original, portanto as análises que seguem baseiam-se na original.

Análise de *Outliers*

Utilizaram-se os filtros *InterquatileRange (IQR)* para se detectar potenciais *outliers* e *SubsetByExpression* a fim de se remover possíveis *outliers.* Abaixo exibem-se os resultados para duas configurações do IQR. Com base na análise exploratória, a análise de outliers se dá sobre o atributo *VALUE24* (o atributo 11).

OBS: *-do-not-check-capabilities* tem que ser marcado para que o filtro possa ser aplicado com todos os atributos. Este problema foi apontado em sala e sugeriu-se remover todos os demais atributos exceto o sob análise. Contudo, deste modo, pode-se manter todos os atributos e avaliar apenas o de interesse.

Naive

InterquartileRange -R 11 -O 3.0 -E 6.0 -do-not-check-capabilities

```
Correctly Classified Instances         197               64.3791 %
Incorrectly Classified Instances       109               35.6209 %
```

InterquartileRange -R 11 -O 1.5 -E 3.0 -do-not-check-capabilities

```
Correctly Classified Instances         192               62.7451 %
Incorrectly Classified Instances       114               37.2549 %
```

KNN (IBk)

InterquartileRange -R 11 -O 3.0 -E 6.0 -do-not-check-capabilities

k = 1

```
Correctly Classified Instances         192               62.7451 %
Incorrectly Classified Instances       114               37.2549 %
```

k = 3

Correctly Classified Instances	186	60.7843 %
Incorrectly Classified Instances	120	39.2157 %

k = 5

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

InterquartileRange -R 11 -O 1.5 -E 3.0 -do-not-check-capabilities

k = 1

Correctly Classified Instances	189	61.7647 %
Incorrectly Classified Instances	117	38.2353 %

k = 3

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

k = 5

Correctly Classified Instances	183	59.8039 %
Incorrectly Classified Instances	123	40.1961 %

PART

InterquartileRange -R 11 -O 3.0 -E 6.0 -do-not-check-capabilities

C = 0.10

Correctly Classified Instances	197	64.3791 %
Incorrectly Classified Instances	109	35.6209 %

C = 0.75

Correctly Classified Instances	189	61.7647 %
Incorrectly Classified Instances	117	38.2353 %

InterquartileRange -R 11 -O 1.5 -E 3.0 -do-not-check-capabilities

C = 0.10

Correctly Classified Instances	189	61.7647 %
Incorrectly Classified Instances	117	38.2353 %

C = 0.75

Correctly Classified Instances	176	57.5163 %
Incorrectly Classified Instances	130	42.4837 %

J48

InterquartileRange -R 11 -O 3.0 -E 6.0 -do-not-check-capabilities

C = 0.10

Correctly Classified Instances	196	64.0523 %
Incorrectly Classified Instances	110	35.9477 %

C = 0.50

Correctly Classified Instances	192	62.7451 %
Incorrectly Classified Instances	114	37.2549 %

InterquartileRange -R 11 -O 1.5 -E 3.0 -do-not-check-capabilities

C = 0.10

Correctly Classified Instances	194	63.3987 %
Incorrectly Classified Instances	112	36.6013 %

C = 0.50

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

Conclusão

Em geral os resultados obtidos pioraram, logo mantém-se a análise sobre a base original.

Análise de *Missing*

Como a base sob análise ainda é a original, aproveitam-se as bases imputadas do primeiro trabalho, sendo elas agora avaliadas nos demais métodos.

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

Naive

```
Correctly Classified Instances         197               64.3791 %
Incorrectly Classified Instances       109               35.6209 %
```

KNN (IBk)

k = 1

```
Correctly Classified Instances         192               62.7451 %
Incorrectly Classified Instances       114               37.2549 %
```

k = 3

```
Correctly Classified Instances         188               61.4379 %
Incorrectly Classified Instances       118               38.5621 %
```

k = 5

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

PART

C = 0.10

```
Correctly Classified Instances         192               62.7451 %
Incorrectly Classified Instances       114               37.2549 %
```

C = 0.75

Correctly Classified Instances	181	59.1503 %
Incorrectly Classified Instances	125	40.8497 %

J48

C = 0.10

Correctly Classified Instances	197	64.3791 %
Incorrectly Classified Instances	109	35.6209 %

C = 0.50

Correctly Classified Instances	196	64.0523 %
Incorrectly Classified Instances	110	35.9477 %

No geral, a remoção de *missing* não produziu modelos melhores, tirando uma pequena melhoria insignificante nos modelos do *J48*, o que condiz que o que fora observado na Avaliação 1. Ainda assim, o melhor resultado observado até agora foi de ~66% para o *PART* com *C=0.1* sobre a base original, logo a remoção de *missing* não nos trouxe muitos ganhos.

Agora apresentam-se dois métodos através dos quais imputaram-se os *missing*. Primeiramente, utiliza-se o método do *Weka, ReplaceMissingValues*, que realiza as substituições através de modas e médias, sem discriminar as instâncias por classe antes de efetuar a imputação.

Naive

Correctly Classified Instances	198	64.7059 %
Incorrectly Classified Instances	108	35.2941 %

KNN (IBk)

k = 1

Correctly Classified Instances	194	63.3987 %
Incorrectly Classified Instances	112	36.6013 %

k = 3

Correctly Classified Instances	189	61.7647 %
Incorrectly Classified Instances	117	38.2353 %

k = 5

Correctly Classified Instances	182	59.4771 %
Incorrectly Classified Instances	124	40.5229 %

PART

C = 0.10

Correctly Classified Instances	192	62.7451 %
Incorrectly Classified Instances	114	37.2549 %

C = 0.75

Correctly Classified Instances	182	59.4771 %
Incorrectly Classified Instances	124	40.5229 %

J48

C = 0.10

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

C = 0.50

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

Este método se saiu pior no geral do que o controle, exceto para o KNN com K=1. 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, os dados foram segmentados de acordo com a classe, para forçar uma imputação relativa a classe. Vejamos como a nova base se saiu:

Naive

```
Correctly Classified Instances         196               64.0523 %
Incorrectly Classified Instances       110               35.9477 %
```

KNN (IBk)

k = 1

```
Correctly Classified Instances         192               62.7451 %
Incorrectly Classified Instances       114               37.2549 %
```

k = 3

```
Correctly Classified Instances         188               61.4379 %
Incorrectly Classified Instances       118               38.5621 %
```

k = 5

```
Correctly Classified Instances         183               59.8039 %
Incorrectly Classified Instances       123               40.1961 %
```

PART

C = 0.10

Correctly Classified Instances	203	66.3399 %
Incorrectly Classified Instances	103	33.6601 %

C = 0.75

Correctly Classified Instances	194	63.3987 %
Incorrectly Classified Instances	112	36.6013 %

J48

C = 0.10

Correctly Classified Instances	189	61.7647 %
Incorrectly Classified Instances	117	38.2353 %

C = 0.50

Correctly Classified Instances	197	64.3791 %
Incorrectly Classified Instances	109	35.6209 %

Este modelo teve desempenho similar ou ligeiramente inferior ao obtido com a base original. Deste modo, continua-se utilizando a base original para as análises posteriores.

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; o script utilizado encontra-se no apêndice.

Naive

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

KNN (IBk)

k = 1

```
Correctly Classified Instances         192               62.7451 %
Incorrectly Classified Instances       114               37.2549 %
```

k = 3

```
Correctly Classified Instances         188               61.4379 %
Incorrectly Classified Instances       118               38.5621 %
```

k = 5

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

PART

C = 0.10

```
Correctly Classified Instances         203               66.3399 %
Incorrectly Classified Instances       103               33.6601 %
```

C = 0.75

Correctly Classified Instances	202	66.0131 %
Incorrectly Classified Instances	104	33.9869 %

J48

C = 0.10

Correctly Classified Instances	196	64.0523 %
Incorrectly Classified Instances	110	35.9477 %

C = 0.50

Correctly Classified Instances	197	64.3791 %
Incorrectly Classified Instances	109	35.6209 %

Conclusão

Obteve-se melhoria significativa para o modelo do *Naive* e os outros demonstraram resultados iguais aos obtidos com a base original. Sendo assim, as análises que seguem baseiam-se na **base de dados orignal** apenas **normalizada**.

Análise de Discretização

Finalmente verifica-se a transformação de dados numéricos em categóricos para averiguar se este processo contribui para melhorias nos resultados. Dos atributos numéricos que temos passíveis de discretização, os principais são *AGE, INCOME, FICO* e *VALUE24.* Realizaram-se e apresentam-se abaixo as mesmas discretizações executadas na Avaliação 1; os scripts utilizados para a produção das discretizações em modo *Batch* encontram-se no Apêndice.

AGE, INCOME, FICO: 5 bins; **VALUE24**: 10 bins: os resultados pioraram para a maior parte dos modelos, portanto a base resultante é descartada.

Naive

```
Correctly Classified Instances         189               61.7647 %
Incorrectly Classified Instances       117               38.2353 %
```

KNN (IBk)

k = 1

```
Correctly Classified Instances         174               56.8627 %
Incorrectly Classified Instances       132               43.1373 %
```

k = 3

```
Correctly Classified Instances         174               56.8627 %
Incorrectly Classified Instances       132               43.1373 %
```

k = 5

```
Correctly Classified Instances         170               55.5556 %
Incorrectly Classified Instances       136               44.4444 %
```

PART

C = 0.10

Correctly Classified Instances	183	59.8039 %
Incorrectly Classified Instances	123	40.1961 %

C = 0.75

Correctly Classified Instances	197	64.3791 %
Incorrectly Classified Instances	109	35.6209 %

J48

C = 0.10

Correctly Classified Instances	198	64.7059 %
Incorrectly Classified Instances	108	35.2941 %

C = 0.50

Correctly Classified Instances	195	63.7255 %
Incorrectly Classified Instances	111	36.2745 %

AGE, INCOME: 5 bins; **FICO, VALUE24**: 6 bins: observa-se uma pequena melhoria no modelo com *C=0.5* da *J48* e o mesmo melhor resultado obtido anteriormente na *PART*, mas de resto não há muitos ganhos.

Naive

Correctly Classified Instances	188	61.4379 %
Incorrectly Classified Instances	118	38.5621 %

KNN (IBk)

k = 1

Correctly Classified Instances	167	54.5752 %
Incorrectly Classified Instances	139	45.4248 %

k = 3

Correctly Classified Instances	175	57.1895 %
Incorrectly Classified Instances	131	42.8105 %

k = 5

Correctly Classified Instances	182	59.4771 %
Incorrectly Classified Instances	124	40.5229 %

PART

C = 0.10

Correctly Classified Instances	203	66.3399 %
Incorrectly Classified Instances	103	33.6601 %

C = 0.75

Correctly Classified Instances	198	64.7059 %
Incorrectly Classified Instances	108	35.2941 %

J48

C = 0.10

Correctly Classified Instances	198	64.7059 %
Incorrectly Classified Instances	108	35.2941 %

C = 0.50

Correctly Classified Instances	199	65.0327 %
Incorrectly Classified Instances	107	34.9673 %

AGE, INCOME, FICO, VALUE24: 10 bins

Naive

Correctly Classified Instances	187	61.1111 %
Incorrectly Classified Instances	119	38.8889 %

KNN (IBk)

k = 1

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

k = 3

Correctly Classified Instances	181	59.1503 %
Incorrectly Classified Instances	125	40.8497 %

k = 5

Correctly Classified Instances	183	59.8039 %
Incorrectly Classified Instances	123	40.1961 %

PART

C = 0.10

Correctly Classified Instances	195	63.7255 %
Incorrectly Classified Instances	111	36.2745 %

C = 0.75

```
Correctly Classified Instances         182               59.4771 %
Incorrectly Classified Instances       124               40.5229 %
```

J48

C = 0.10

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

C = 0.50

```
Correctly Classified Instances         199               65.0327 %
Incorrectly Classified Instances       107               34.9673 %
```

Conclusão

Uma observação interessante que se contata nesta seção é a predominância do *PART* em apresentar bons resultados, desde o início do trabalho, para C=0.1. Exceto pela melhoria considerável que se constatou no Naive quando da simples normalização da base, o PART dominou os resultados.

Em razão disto, adota-se a base de dados normalizada com a segunda discretização apresentada acima e o modelo *PART* com *C = 0.1*, modelo este que obteve o melhor resultado dentre todos os demais (66.3399%). Embora este resultado já tivesse surgido na análise da base original, agora temos uma base mais simples, devido aos processos de normalização e, principalmente, discretização, o qual empacota em um número reduzido de classes a mesma informação antes passada por um domínio real.

Ainda é válido observar que a discretização contribuiu para melhorias nos resultados obtidos pelo *J48*, o que fora observado no primeiro trabalho.

Seleção de variáveis

Utilizando-se a base de dados pré-processada definida na conclusão da seção anterior, apresenta-se o resultado do processo de seleção de variáveis. Os resultados completos encontram-se anexados ao Apêndice.

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

É nítido que há um consenso que os atributos *LOCATION, INCOME, BUY18* e *VALUE24* são de extrema relevância. Por outro lado, *BUY6, BUY12*, *GENDER* e *FICO* figuram frequentemente no final do *rank*. Avaliemos agora o resultado de eliminar tais atributos. Em anexo, no apêndice, disponibilizam-se os *scripts* usados para a remoção dos atributos.

Sem *BUY6*

Correctly Classified Instances	190	62.0915 %
Incorrectly Classified Instances	116	37.9085 %

Sem *BUY12*

Correctly Classified Instances	188	61.4379 %
Incorrectly Classified Instances	118	38.5621 %

Sem *GENDER*

Correctly Classified Instances	181	59.1503 %
Incorrectly Classified Instances	125	40.8497 %

Sem *FICO*

Correctly Classified Instances	190	62.0915 %
Incorrectly Classified Instances	116	37.9085 %

Sem Todos

Correctly Classified Instances	190	62.0915 %
Incorrectly Classified Instances	116	37.9085 %

Conclusão

Apesar dos métodos de seleção de variáveis indicarem o mesmo grupo de atributos como os menos relevantes, sua remoção resulta em uma piora do modelo, o que indica que o *PART* os está usando para a classificação; a remoção do *GENDER* inclusive acarreta uma queda de aproximadamente 7 pontos percentuais na taxa de acerto! Portanto, não é possível simplificar a base através da remoção de atributos.

Conclusão

Em conclusão, fui capaz de obter um modelo com taxa de acerto de 66.3399% com o modelo *PART* para *C = 0.1*, executado sobre uma base resultante da normalização da base original e da discretização de duas variáveis numéricas contínuas (*FICO* e *VALUE24*) em 6 *bins* e de outras duas (*AGE* e *INCOME*) em 5 *bins*.

Infelizmente não houve melhorias nos resultados obtidos com este modelo, que desde a base original apresentou o mesmo resultado. Por outro lado, a base de dados ficou mais "enxuta", visto que o processo de normalização e discretização simplificou alguns de seus atributos.

Apêndice

Scripts

Normalização

```
java weka.filters.unsupervised.attribute.Standardize -b -i
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buytreinamento_sem_preproc
essamento_fixed_types.arff -o
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buy_treinamento_std.arff
-r
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_sem_preprocessame
nto_fixed_types.arff -s
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_std.arff
```

Discretização

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

```
# Discretize VALUE24 to 10 bins
java weka.filters.unsupervised.attribute.Discretize -B 10 -M -1.0 -R 11 -b -i
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buy_treinamento_std.arff
-o
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buy_treinamento_std_value2
4b10.arff -r
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_std.arff -s
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_std_value24b10.ar
ff

# Discretize AGE, INCOME, FICO to 5 bins
java weka.filters.unsupervised.attribute.Discretize -B 5 -M -1.0 -R 1,2,5 -b -i
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buy_treinamento_std_value2
4b10.arff -o
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buy_treinamento_std_value2
4b10_ageIncomeFicoB5.arff -r
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_std_value24b10.ar
ff -s
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_std_value24b10_ag
eIncomeFicoB5.arff
```

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

```
# Discretize VALUE24, FICO to 6 bins
java weka.filters.unsupervised.attribute.Discretize -B 6 -M -1.0 -R 5,11 -b -i
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buy_treinamento_std.arff
-o
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buy_treinamento_std_value2
4FicoB6.arff -r
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_std.arff -s
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_std_value24FicoB6
.arff
```

```
# Discretize AGE, INCOME to 5 bins
java weka.filters.unsupervised.attribute.Discretize -B 5 -M -1.0 -R 1,2 -b -i
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buy_treinamento_std_value2
4FicoB6.arff -o
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buy_treinamento_std_value2
4FicoB6_ageIncomeB5.arff -r
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_std_value24FicoB6
.arff -s
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_std_value24FicoB6
_ageIncomeB5.arff
```

AGE, INCOME, FICO, VALUE24: 10 bins

```
# Discretize AGE, INCOME, FICO, VALUE24 to 10 bins
java weka.filters.unsupervised.attribute.Discretize -B 10 -M -1.0 -R 1,2,5,11 -b -i
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buy_treinamento_std.arff
-o
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buy_treinamento_std_value2
4AgeIncomeFicoB10.arff -r
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_std.arff -s
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_std_value24AgeInc
omeFicoB10.arff
```

Remoção de Atributos

```
# Remove BUY6
java weka.filters.unsupervised.attribute.Remove -R 8 -b -i
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buy_treinamento_std_value2
4FicoB6_ageIncomeB5.arff -o
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buy_treinamento_std_value2
4FicoB6_ageIncomeB5_semBuy6.arff -r
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_std_value24FicoB6
_ageIncomeB5.arff -s
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_std_value24FicoB6
_ageIncomeB5_semBuy6.arff

# Remove BUY12
java weka.filters.unsupervised.attribute.Remove -R 9 -b -i
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buy_treinamento_std_value2
4FicoB6_ageIncomeB5.arff -o
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buy_treinamento_std_value2
4FicoB6_ageIncomeB5_semBuy12.arff -r
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_std_value24FicoB6
_ageIncomeB5.arff -s
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_std_value24FicoB6
_ageIncomeB5_semBuy12.arff

# Remove GENDER
java weka.filters.unsupervised.attribute.Remove -R 3 -b -i
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buy_treinamento_std_value2
4FicoB6_ageIncomeB5.arff -o
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buy_treinamento_std_value2
4FicoB6_ageIncomeB5_semGender.arff -r
```

```
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_std_value24FicoB6
_ageIncomeB5.arff -s
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_std_value24FicoB6
_ageIncomeB5_semGender.arff

# Remove FICO
java weka.filters.unsupervised.attribute.Remove -R 5 -b -i
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buy_treinamento_std_value2
4FicoB6_ageIncomeB5.arff -o
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buy_treinamento_std_value2
4FicoB6_ageIncomeB5_semFico.arff -r
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_std_value24FicoB6
_ageIncomeB5.arff -s
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_std_value24FicoB6
_ageIncomeB5_semFico.arff

# Remove All
java weka.filters.unsupervised.attribute.Remove -R 3,5,8,9 -b -i
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buy_treinamento_std_value2
4FicoB6_ageIncomeB5.arff -o
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buy_treinamento_std_value2
4FicoB6_ageIncomeB5_semTodos.arff -r
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_std_value24FicoB6
_ageIncomeB5.arff -s
/Users/calmattoso/OneDrive/PUC/ENG1034_DataMining/Trabalho_2/buyteste_std_value24FicoB6
_ageIncomeB5_semTodos.arff
```

Resultados de Seleção de Variáveis

OneR

```
Attribute Evaluator (supervised, Class (nominal): 13 class):
       OneR feature evaluator.

       Using 10 fold cross validation for evaluating attributes.
       Minimum bucket size for OneR: 6

Ranked attributes:
78.176   7 LOCATION
60.342   2 INCOME
56.678  11 VALUE24
56.352  10 BUY18
55.537   1 AGE
52.85    4 MARRIED
52.606   5 FICO
51.71    3 GENDER
51.71    6 OWNHOME
51.629   9 BUY12
51.221  12 MOVED
50.814   8 BUY6

Selected attributes: 7,2,11,10,1,4,5,3,6,9,12,8 : 12
```

Gain

```
Attribute Evaluator (supervised, Class (nominal): 13 class):
Gain Ratio feature evaluator

Ranked attributes:
 0.137036   7 LOCATION
 0.016712   2 INCOME
 0.016158  12 MOVED
 0.014256  10 BUY18
 0.009607  11 VALUE24
 0.005798   1 AGE
 0.001938   6 OWNHOME
 0.00165    5 FICO
 0.001458   4 MARRIED
 0.000665   3 GENDER
 0          9 BUY12
 0          8 BUY6

Selected attributes: 7,2,12,10,11,1,6,5,4,3,9,8 : 12
```

InfoGain

```
Attribute Evaluator (supervised, Class (nominal): 13 class):
Information Gain Ranking Filter

Ranked attributes:
 0.359999   7 LOCATION
 0.033953   2 INCOME
 0.018308  11 VALUE24
 0.013399  10 BUY18
 0.011107   1 AGE
 0.003411  12 MOVED
 0.003139   5 FICO
 0.001637   6 OWNHOME
 0.001406   4 MARRIED
 0.000665   3 GENDER
 0          9 BUY12
 0          8 BUY6

Selected attributes: 7,2,11,10,1,12,5,6,4,3,9,8 : 12
```

Relief

```
Attribute Evaluator (supervised, Class (nominal): 13 class):
ReliefF Ranking Filter
Instances sampled: all
Number of nearest neighbours (k): 10
Equal influence nearest neighbours

Ranked attributes:
```

```
 0.3176710097720034          7 LOCATION
 0.014196524359103372       10 BUY18
 0.012019543973941341        2 INCOME
 0.00838762214983712        11 VALUE24
 0.007459283387622137        1 AGE
 0.005130293159609118        6 OWNHOME
 0.00004071640004246355      9 BUY12
 0.000000000000000006103     3 GENDER
-0.000000000000000000226    12 MOVED
-0.00118078159223427         8 BUY6
-0.0061889250814332174       4 MARRIED
-0.013721498371335497        5 FICO

Selected attributes: 7,10,2,11,1,6,9,3,12,8,4,5 : 12
```

Correlation

```
Attribute Evaluator (supervised, Class (nominal): 13 class):
Correlation Ranking Filter
Ranked attributes:
 0.2819   7 LOCATION
 0.1505  10 BUY18
 0.0991   2 INCOME
 0.0874  11 VALUE24
 0.068   12 MOVED
 0.0592   6 OWNHOME
 0.0589   4 MARRIED
 0.0336   1 AGE
 0.0323   9 BUY12
 0.0311   5 FICO
 0.0137   8 BUY6
 0.0131   3 GENDER

Selected attributes: 7,10,2,11,12,6,4,1,9,5,8,3 : 12
```