Trabalho 1 – Aprendizado Automático

A partir da base de dados enviada e do estudo sobre pré-processamentos, apresentar os resultados obtidos, utilizando o modelo Decision Tree (DT), e discutir seus resultados dentro dos processos para Data Mining.

1) A partir da base de dados:

a. Descreva a base de dados descrendo os atributos, numéricos e categóricos, classificando segundo a escala (nominal ou razão) e a cardinalidade (discreta,

contínua, binária).

|  |  |  |  |
| --- | --- | --- | --- |
| **Atributo** | **Tipo de Dado** | **Escala** | **Cardinalidade** |
| **Age** | Numérico | Razão | Contínua |
| **Sex** | Categórico | Nominal | Binária |
| **Job** | Numérico | Razão | Discreta |
| **Housing** | Categórico | Nominal | Discreta |
| **Saving accounts** | Categórico | Nominal | Discreta |
| **Checking account** | Categórico | Nominal | Discreta |
| **Credit amount** | Numérico | Razão | Contínua |
| **Duration** | Numérico | Razão | Contínua |
| **Purpose** | Categórico | Nominal | Discreta |

b. Descreva cada um dos atributos segundo frequência, mínimo e máximo valor,

dia desvios padrão, conforme o caso.

Atributos Numéricos:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Atributo** | **Mín.** | **25%** | **Mediana** | **75%** | **Máx.** | **Média** | **Desvio Padrão** |
| **Age** | 19 | 27 | 33 | 42 | 75 | 35,55 | 11,38 |
| **Job** | 0 | 2 | 2 | 2 | 3 | 1,90 | 0,65 |
| **Credit amount** | 250 | 1365 | 2319 | 3972 | 18424 | 3271,26 | 2822,74 |
| **Duration** | 4 | 12 | 18 | 24 | 72 | 20,90 | 12,06 |

Podemos ver que o atributo Credit Amount apresenta alta variabilidade, tendo o maior desvio padrão e assimétrico, pois a média e a mediana estão distantes, assim podemos acreditar que tem outliers. Duration apresenta também mediana e média distantes. O Age está relativamente concentrado. O job está entre 0 e 3, com maior concentração na classificação 2

Atributos Categóricos:

\* Unknown foi usado para preencher os itens em branco

- **Sex**:

* male: 690
* female: 310

- **Housing**:

* own: 713
* rent: 179
* free: 108

- **Saving accounts**:

* little: 603
* unknown \*: 183
* moderate: 103
* quite rich: 63
* rich: 48

- **Checking account**:

* unknown \*: 394
* little: 274
* moderate: 269
* rich: 63

- **Purpose**:

* car: 337
* radio/TV: 280
* furniture/equipment: 181
* business: 97
* education: 59
* repairs: 22
* domestic appliances: 12
* vacation/others: 12

A maior parte dos clientes possui conta de poupança classificada como little e conta corrente como unknown. O propósito mais comum do crédito é a compra de carro ou eletrodomésticos.

O atributo Risk, classe alvo, apresenta leve desbalanceamento, com 700 exemplos "good" e 300 "bad".

c. Avalie os resultados dos processos abaixo, caso sejam utilizados na base de

dados, após o processo de classificação com DT ter sido utilizado.

Modelo Base apresentou:

* Acurácia 0.66
* Recall “bad”: 0.48
* Recall “good”: 0.74

Conclusão: Modelo inicial mostra viés para a classe majoritária (“good”). Baixo desempenho para “bad”.

1. limpeza de dados (outlier, missing)

**Resultado**:

* Acurácia aumentou de 0.66 (modelo base) para 0.71
* Recall da classe "bad" subiu de 0.48 para 0.51
* Recall da classe "good" subiu de 0.74 para 0.80

Conclusão: Tratamento de valores ausentes melhora o equilíbrio e acurácia geral..

1. normalização/transformação

**Resultado obtido**:

* Acurácia manteve-se em 0.71
* Recall para "bad" = 0.52, e para "good" = 0.79

Conclusão: Sem efeito prático em DT, como esperado (Decision Tree é insensível à escala).

1. discretização

**Resultado**:

* Acurácia caiu de 0.66 (modelo base) para 0.63
* Recall da classe "bad" caiu de 0.48 para 0.43
* Recall da classe "good" também caiu levemente para 0.71

Conclusão: Discretização reduziu desempenho. Binning não ajudou neste cenário.

Portanto, podemos concluir que a limpeza de dados foi o único pré-processamento que melhorou significativamente a performance da árvore, tanto em acurácia quanto em equilíbrio entre as classes.  
A normalização teve efeito neutro, como esperado para árvores.  
A discretização reduziu o desempenho, indicando que os bins utilizados não capturaram bem as diferenças entre as classes.

2) Faça uma análise de seleção de variáveis utilizando os métodos que julgar necessários, considerando o tipo de dado. Abaixo uma tabela é indicada para que possa usar como inspiração avaliação de importância composta dos métodos que considerar.

Dica: utilize uma tabela métodos x ranking com atributos selecionados

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | Método de Seleção 1 | Método de Seleção 1 | Método de Seleção 1 | Método de Seleção 1 |  | | Atributo 1 |  |  |  |  |  | | Atributo 2 |  |  |  |  |  | | ... |  |  |  |  |  | | Atributo n |  |  |  |  |  | | | | |
|  | | | |
|  | | | |
| Os métodos de seleção usados:  1. Chi², avalia a associação estatística entre cada atributo e a variável alvo.  2. Correlação de Pearson, mede a força da relação linear entre variáveis contínuas e a classe (convertida em binária).  3. Importância da Árvore, usa uma árvore de decisão para calcular a contribuição de cada atributo nos splits.  4. RFE (Recursive Feature Elimination) com árvore, elimina atributos menos relevantes com base na performance da árvore.   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Atributo** | **Chi²** | **Pearson** | **Importância Árvore** | **Ranking RFE** | | Credit amount | 58262.49 | 0.1547 | 0.2796 | 1 | | Age | 30.20 | 0.0911 | 0.1517 | 2 | | Duration | 321.03 | 0.2149 | 0.1340 | 3 | | Checking account | 123.33 | 0.3508 | 0.1496 | 4 | | Purpose | 5.08 | 0.0611 | 0.1218 | 5 | | Saving accounts | 72.27 | 0.1789 | 0.0554 | 6 | | Job | 0.24 | 0.0327 | 0.0491 | 7 | | Housing | 0.10 | 0.0193 | 0.0377 | 8 | | Sex | 1.77 | 0.0755 | 0.0212 | 9 |   Com base nos resultados acima, selecionei os cinco atributos mais relevantes de forma consistente entre os métodos que foram: Credit amount, Age, Duration, Checking account, Purpose.  Treinei um novo modelo de árvore de decisão utilizando apenas esse subconjunto, e obtive uma acurácia de 0.703. Este valor é muito próximo da acurácia máxima obtida com a base completa (0.71), apesar da acurácia não ter aumentado, o modelo ficou mais eficiente, interpretável e robusto. | | | |