Trabalho de IA

Christian Franchin dos Santos
Felipe da Maia Bueno

Introdução

Trabalho referente ao conteúdo de mineração de dados.

Para o estudo de caso escolhemos uma sub-base da Pesquisa Nacional De Prevalência De Métodos Contraceptivos Da Indonésia, realizada em 1987. Pesquisa essa que foi realizada com mulheres casadas.

A ferramenta WEKA foi utilizada.

Introdução

A base possui dados coletados 1473 mulheres.

A classificação foi feita da seguinte forma:

- Não utiliza (629);
- Utiliza de curto período (333);
- Utiliza de longo período (511);

Domínio da Aplicação

Questões Religiosas

- É permitido utilizar toda e qualquer forma de método contraceptivo;
- Se uma gravidez puder causar perigo a vida de uma mulher, esta pode adotar métodos contraceptivos;
- Tanto a Tubectomia e Vasectomia são permitidas, desde que o procedimento seja reversível;
- A mulher não pode utilizar métodos contraceptivos sem a aprovação de seu marido.

Atributos

Idade da Mulher	-
Educação da Mulher	Baixa, Média Baixa, Média Alto , Alto
Educação do Marido	Baixa, Média Baixa, Média Alto , Alto
Número de Filhos	-
Religião da Mulher	Islâmica, Não Islâmica
Mulher está trabalhando?	Sim, Não
Ocupação do Marido	Baixa, Média Baixa, Média Alto , Alto
Padrão de Vida	Baixa, Média Baixa, Média Alto , Alto
Exposição à Mídia	Boa, Não Boa
Método Contraceptivo Utilizado	Nenhum, Curto Prazo, Longo Prazo

42.70%

Erro da classe majoritária (Não utiliza método contraceptivo)

Métodos aplicados

Cross-Validation

É uma técnica para medir como os resultados de uma análise estatística vão ser generalizados para um conjunto de dados independente.

Ela é principalmente usada onde o objetivo é a predição, e se deseja estimar o quão correto um modelo preditivo irá ser executado na prática.

Foram utilizados 10 folds.

Algoritmo C4.5

O algoritmo C4.5 é utilizado para a geração de árvores de decisão e utiliza a lógica de dividir e conquistar.

Funcionamento:

- Caso o nó atual possua um alto conjunto da mesma classe ou não tenha nenhum ganho de informação, essa é definida como nó folha.
- Do contrário um atributo é selecionado para particionar o conjunto em subconjuntos e o processo é aplicado recursivamente.

Algoritmo C4.5

- Lida com atributos categóricos;
- Faz uma busca na árvore, de baixo para cima, e transforma em nós folha aqueles ramos que não apresentam nenhum ganho significativo;
- Utiliza 2 heurísticas: Ganho de informação e a Taxa de Ganho (ganho de informação / informações fornecidas);
- Árvore gerada é de fácil compreensão.

C4.5

Classificados corretamente	52.1385 %
Estatística kappa	0.2549
Erro médio absoluto	0.356
Erro quadrático médio da raiz	0.4736
Erro relativo absoluto	82.6287 %
Erro quadrático relativo a raiz	102.0514 %

```
=== Confusion Matrix ===
  a   b   c   <-- classified as
383  61 185 |   a = No-Use
 86 126 121 |   b = Long-Term
156  96 259 |   c = Short-Term
```

Multilayer Perceptron

- Rede Neural com várias estruturas combinadas, onde cada uma delas possui um número diferente de neurônios.

Funcionamento:

- A rede neural foi construída automaticamente, com valores padrão;
- Composta por (atributos + classes)/2 com taxa de aprendizado de 0.3 o momentum 0.2;

Multilayer Perceptron

- Flexíveis e podem ser usadas para problemas de regressão e classificação;
- Uma vez treinados, as previsões são bem rápidas;
- É confiável em uma abordagem de tarefas que envolvem muitos recursos. Funciona dividindo o problema da classificação em uma rede em camadas de elementos mais simples;

Multilayer Perceptron

Classificados corretamente	52.3422%
Estatística kappa	0.2599
Erro médio absoluto	0.3434
Erro quadrático médio da raiz	0.4633
Erro relativo absoluto	79.7012%
Erro quadrático relativo a raiz	99.8144%

```
=== Confusion Matrix ===
   a   b   c   <-- classified as
 387  70 172  |   a = No-Use
  97 135 101  |   b = Long-Term
 155 107 249 |   c = Short-Term
```

Seleção de Atributos por Rank

Ganho de informação

1. Children
2. W.Education
3. W.Age
4. H.Education

---

5. Standard-of-living
6. H.occupation
7. Media Exposure
8. W.Islamic
9. W.Working

C4.5 com seleção atributos

Classificados corretamente	53.4963 %
Estatística kappa	0.2793
Erro médio absoluto	0.3524
Erro quadrático médio da raiz	0.4449
Erro relativo absoluto	81.8035 %
Erro quadrático relativo a raiz	95.8552 %

```
=== Confusion Matrix ===
   a   b   c   <-- classified as
 385  59 185 |   a = 1
  67 131 135 |   b = 2
 135 104 272 |   c = 3
```

Multilayer Perceptron com seleção atributos

Classificados corretamente	55.7366 %
Estatística kappa	0.3148
Erro médio absoluto	0.3507
Erro quadrático médio da raiz	0.4281
Erro relativo absoluto	81.3932 %
Erro quadrático relativo a raiz	92.2345 %

```
=== Confusion Matrix ===
   a   b   c   <-- classified as
 389  63 177 |   a = 1
  73 136 124 |   b = 2
 117  98 296 |   c = 3
```

Considerações

Resultados

	Sem Seleção de Atributos	Com Seleção de Atributos
C4.5	52.1385 %	53.4963 %
Multilayer Perceptron	52.3422 %	**55.7366 %**

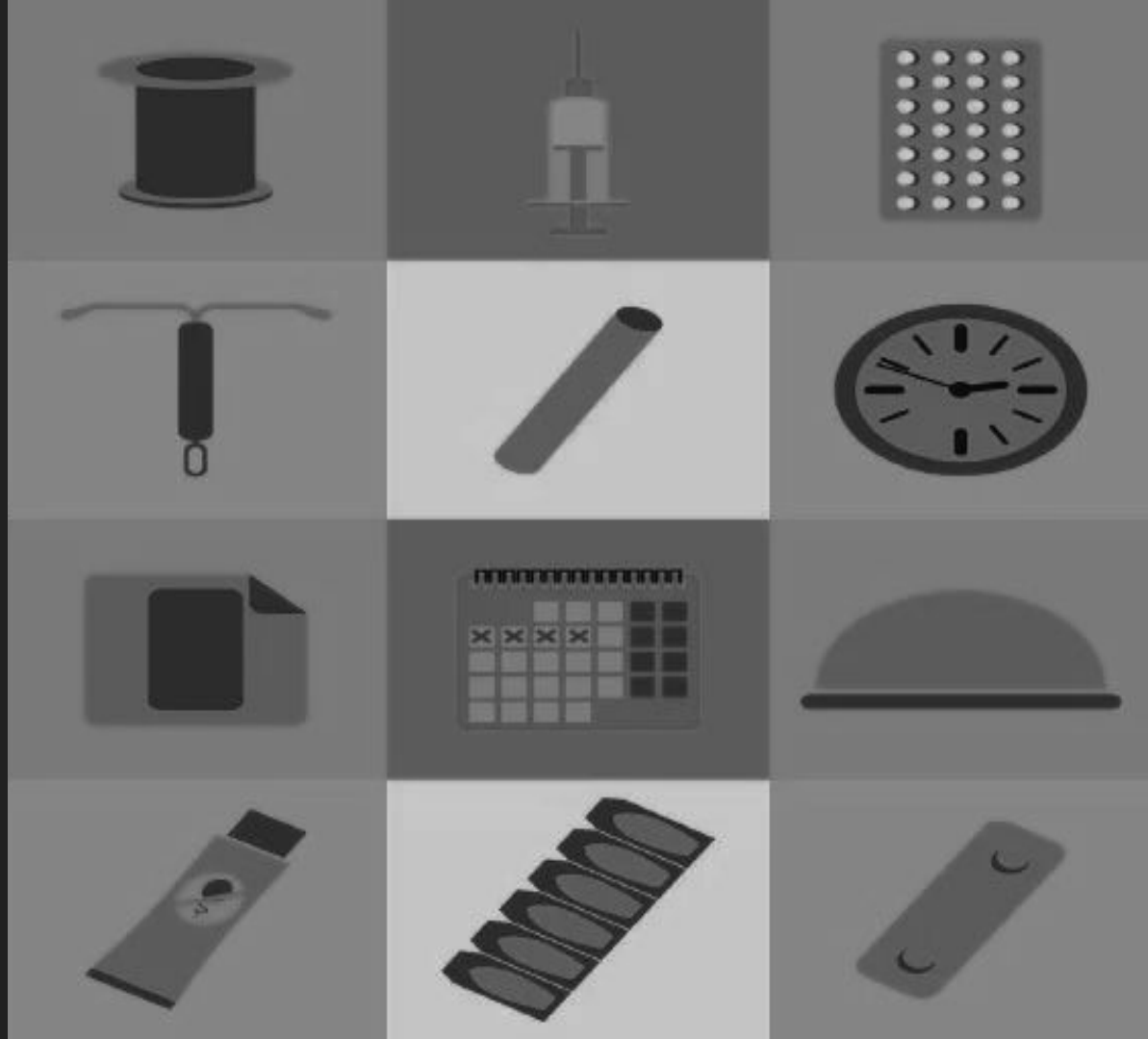
Comparação

Algoritmo	Contraceptive Method
0-R	42,7020%
SMO	50,4413%
Multilayer Perceptron	52,3422%
NaïveBayes	50,7807%
C&RT	55,1935%
Melhor da literatura	**69,7600%**

Obrigado!

Christian Franchin dos Santos
christianfranchin@gmail.com

Felipe da Maia Bueno
felipe_mbsti@hotmail.com



Referências Bibliográficas

- Birth Control. Al-Islam. https://www.al-islam.org/islamic-edicts-on-family-planning/birth-control.
- Top 10 algorithms in data mining. http://www.cs.umd.edu/~samir/498/10Algorithms-08.pdf.
- Estudo sobre Algoritmos de Classificação para o Referenciamento de Gestantes de Alto-risco. ftp://ftp.inf.puc-rio.br/pub/docs/techreports/09_38_nunes.pdf.