Regras de Associação e Métodos de Classificação

André F. B. Menezes

Universidade Estadual de Maringá

Departamento de Estatística

19 de Setembro de 2017

Organização

Introdução

Motivação e Objetivos

- Trabalho para a disciplina Tópicos Especiais em Estatística;
- Aplicar as técnicas de regras de associação e classificação em algum problema real:
  1. Regras de associação: variáveis climatológicas;
  2. Classificação: variáveis climatológicas e partidas de tenis.

Regras de Associação

Conceitos

- Uma regra de associação é uma implicação da forma $A \Rightarrow B$, onde A (antecedente) e B (consequente) são conjuntos disjuntos de itens.
- Medidas importantes:

$$\begin{aligned} \text{support}\,(A \Rightarrow B) &= P\,(A \cap B)\,, \\ \text{confidence}\,(A \Rightarrow B) &= \frac{P\,(A \cap B)}{P\,(A)}, \\ \text{lift}\,(A \Rightarrow B) &= \frac{P\,(A \cap B)}{P\,(A)\ P\,(B)}. \end{aligned}$$

Regras de Associação

Conceitos

- Algoritmo APRIORI, permite selecionar as regras tais que satisfazem um limite inferior para o suporte e confiança (AGRAWAL e SRIKANT, 1994).
- Disponível na função `apriori` da biblioteca `arules`.
- Outras medidas utilizadas:

$$\text{conviction}\,(A \Rightarrow B) = \frac{P(A)\ P\left(\bar{B}\right)}{P\left(A \cap \bar{B}\right)}$$

$$\text{OR}\,(A \Rightarrow B) = \frac{P(A \cap B)\ P\left(\bar{A} \cap \bar{B}\right)}{P\left(A \cap \bar{B}\right)\ P\left(\bar{A} \cap B\right)}$$

- 1 indica independência.

Aplicação I

Dados sobre o Clima de Maringá

- Descrição: 16.062 observações relacionada a variáveis climatológicas da estação de Maringá durante os anos de 1961 a 2016.
- Fonte: Instituo Nacional de Meteorologia (INMET) `http://www.inmet.gov.br/`.
- Objetivo: Determinar relações entre as variáveis e a ocorrência de chuva.

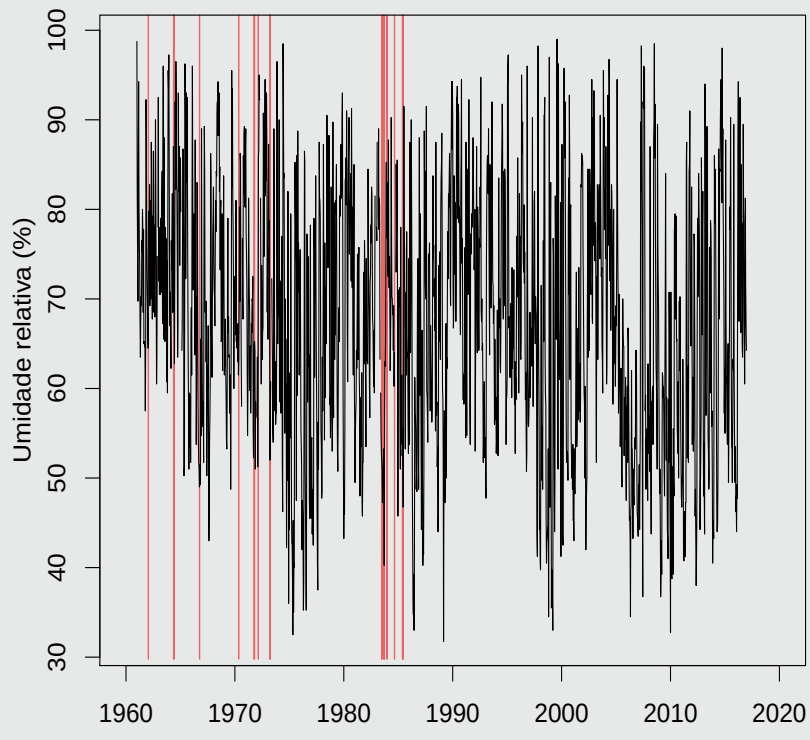
Aplicação I

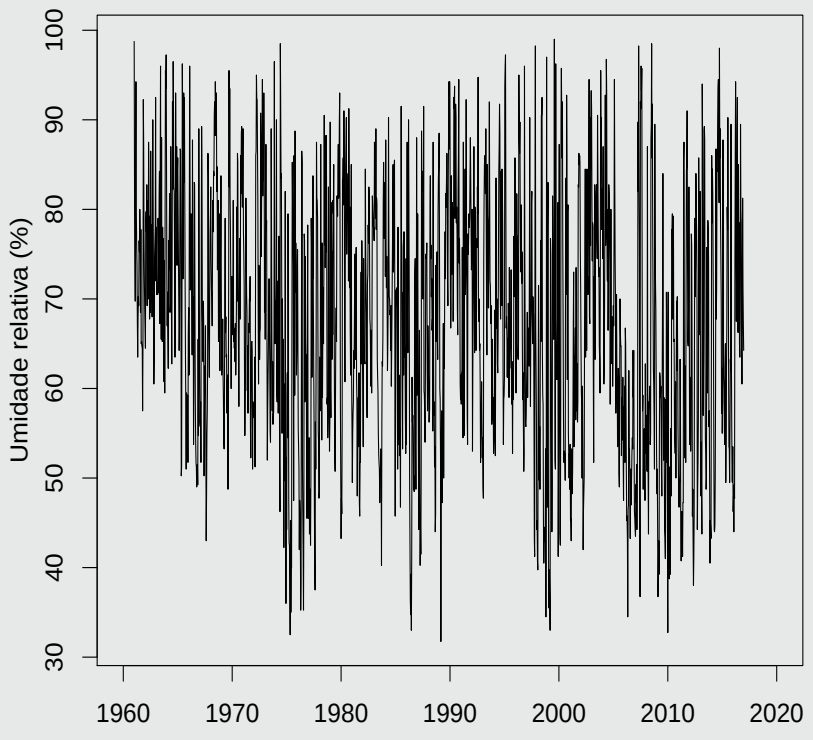
Medidas descritivas das variáveis

Variável	Faltantes	Mín.	Média	Mediana	Máx.
Precipitação (mm)	61	0.00	4.60	0.00	170.30
Umidade Relativa (%)	71	24.00	69.21	69.00	100.00
Velocidade do Vento (mps)	0	0.00	1.41	1.00	8.33
Temperatura Máxima (°C)	137	8.80	28.10	28.70	39.40
Temperatura Mínima (°C)	46	-0.20	17.48	18.20	26.40

Aplicação I

Interpolação das observações faltantes – Umidade relativa



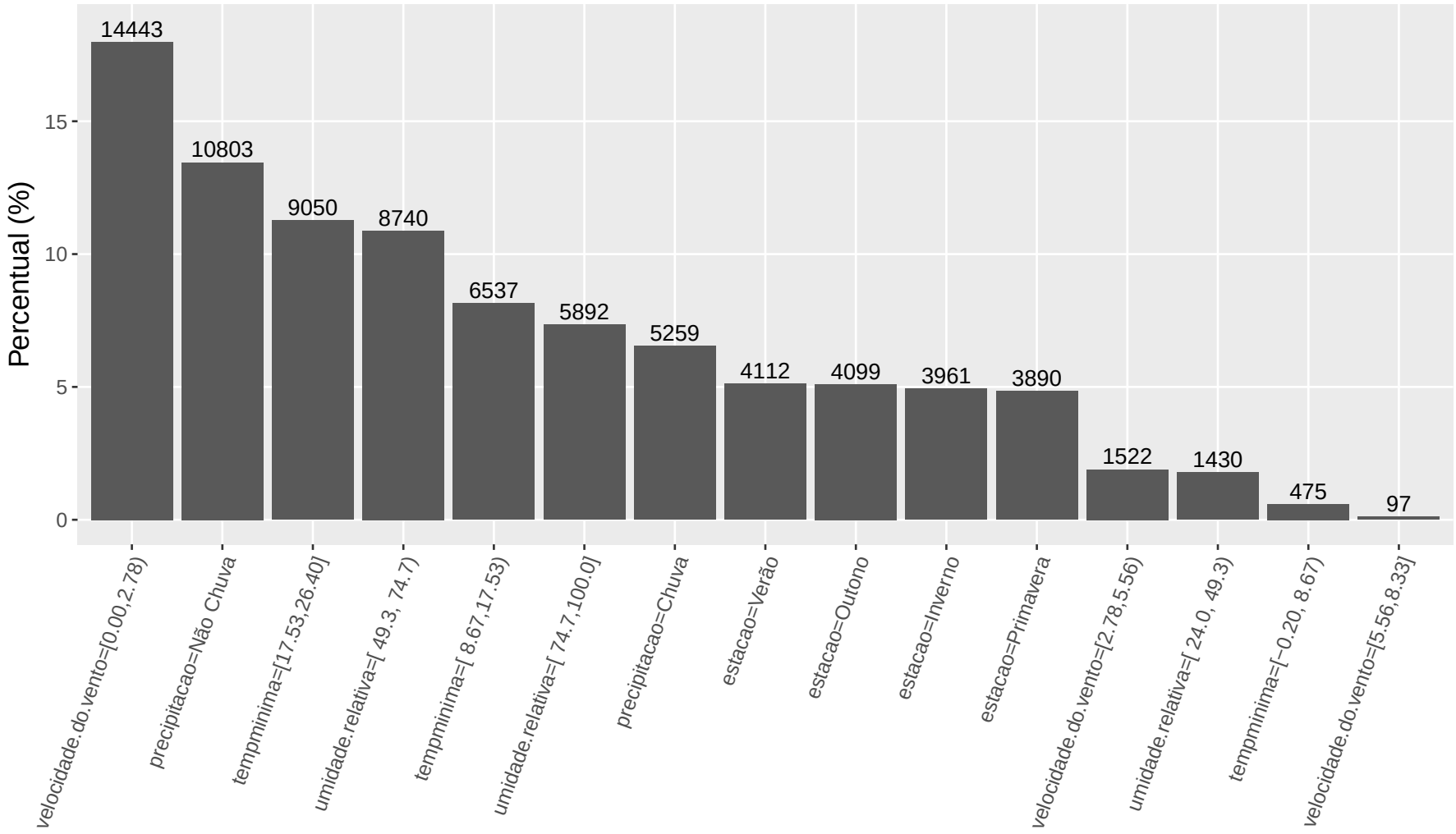


Aplicação I

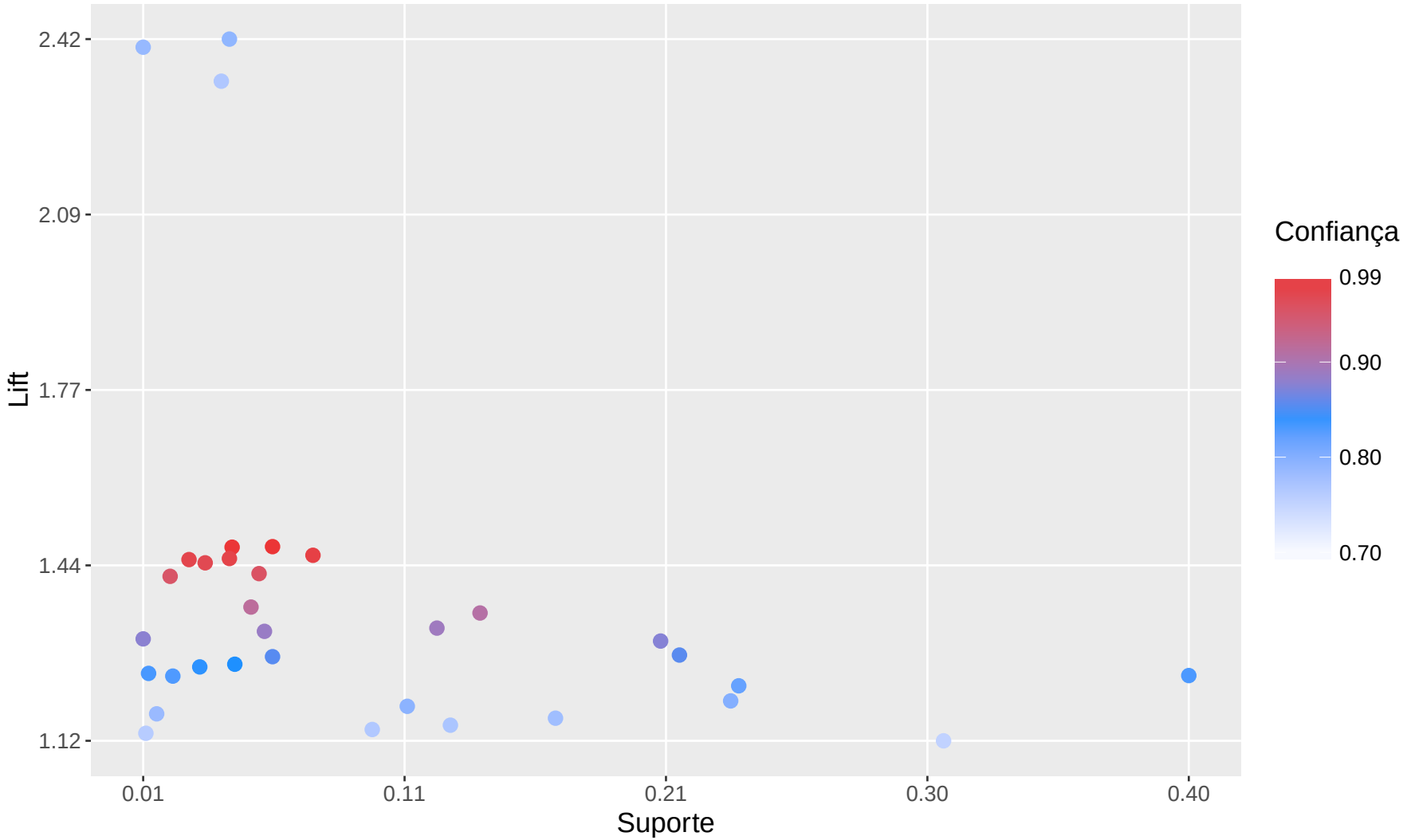
Frequência absoluta (percentual) das variáveis categorizadas

Variável	Categorias	Frequência (%)
Precipitação	Chuva	5259 (32.742)
	Não Chuva	10803 (67.258)
Estação do ano	Inverno	3961 (24.661)
	Outono	4099 (25.520)
	Primavera	3890 (24.219)
	Verão	4112 (25.601)
Umidade relativa (%)	[24.0, 49.3)	1430 (8.903)
	[49.3, 74.7)	8740 (54.414)
	[74.7,100.0]	5892 (36.683)
Velocidade do vento (mps)	[0.00,2.78)	14443 (89.920)
	[2.78,5.56)	1522 (9.476)
	[5.56,8.33]	97 (0.604)
Temperatura máxima (°C)	[8.8,19.0)	508 (3.163)
	[19.0,29.2)	8212 (51.127)
	[29.2,39.4]	7342 (45.710)
Temperatura mínima (°C)	[-0.20, 8.67)	475 (2.957)
	[8.67,17.53)	6537 (40.699)
	[17.53,26.40]	9050 (56.344)

Aplicação I


Percentual (%)
0
5
10
15
14443
10803
9050
8740
6537
5892
5259
4112
4099
3961
3890
1522
1430
475
97
velocidade.do.vento=[0.00,2.78)
precipitacao=Não Chuva
tempminima=[17.53,26.40]
umidade.relativa=[49.3, 74.7)
tempminima=[8.67,17.53)
umidade.relativa=[74.7,100.0]
precipitacao=Chuva
estacao=Verão
estacao=Outono
estacao=Inverno
estacao=Primavera
velocidade.do.vento=[2.78,5.56)
umidade.relativa=[24.0, 49.3)
tempminima=[−0.20, 8.67)
velocidade.do.vento=[5.56,8.33]

Aplicação I


2.42
2.09
1.77
1.44
1.12
Lift
0.01
0.11
0.21
0.30
0.40
Suporte
Confiança
0.99
0.90
0.80
0.70

Aplicação I

Regra ($A \Rightarrow B$)			Suporte	Confiança	Lift	Conviction	OR
tempmaxima=[29.2,39.4],umidade.relativa=[24.0, 49.3)	$\Rightarrow$	precipitacao=Não Chuva	0.062	0.993	1.476	47.195	76.705
umidade.relativa=[24.0, 49.3),tempminima=[17.53,26.40]	$\Rightarrow$	precipitacao=Não Chuva	0.047	0.992	1.475	41.582	65.878
umidade.relativa=[24.0, 49.3),velocidade.do.vento=[0.00,2.78)	$\Rightarrow$	precipitacao=Não Chuva	0.077	0.982	1.460	17.965	29.492
umidade.relativa=[24.0, 49.3),estacao=Inverno	$\Rightarrow$	precipitacao=Não Chuva	0.046	0.978	1.454	14.695	22.873
umidade.relativa=[24.0, 49.3),estacao=Primavera	$\Rightarrow$	precipitacao=Não Chuva	0.031	0.976	1.452	13.861	21.042
umidade.relativa=[24.0, 49.3),tempminima=[8.67,17.53)	$\Rightarrow$	precipitacao=Não Chuva	0.037	0.972	1.446	11.826	18.037
tempmaxima=[29.2,39.4],estacao=Inverno	$\Rightarrow$	precipitacao=Não Chuva	0.057	0.959	1.426	8.060	12.474
tempmaxima=[19.0,29.2),umidade.relativa=[24.0, 49.3)	$\Rightarrow$	precipitacao=Não Chuva	0.024	0.956	1.421	7.367	10.818
tempminima=[17.53,26.40],estacao=Inverno	$\Rightarrow$	precipitacao=Não Chuva	0.054	0.917	1.364	3.958	5.781
umidade.relativa=[49.3, 74.7),estacao=Outono	$\Rightarrow$	precipitacao=Não Chuva	0.139	0.910	1.353	3.640	5.950
umidade.relativa=[49.3, 74.7),estacao=Inverno	$\Rightarrow$	precipitacao=Não Chuva	0.123	0.891	1.325	3.017	4.673
tempmaxima=[29.2,39.4],tempminima=[8.67,17.53)	$\Rightarrow$	precipitacao=Não Chuva	0.059	0.887	1.319	2.909	4.110
tempmaxima=[19.0,29.2),tempminima=[-0.20, 8.67)	$\Rightarrow$	precipitacao=Não Chuva	0.014	0.878	1.305	2.681	3.553
umidade.relativa=[49.3, 74.7),tempminima=[8.67,17.53)	$\Rightarrow$	precipitacao=Não Chuva	0.206	0.875	1.301	2.615	4.458
tempmaxima=[19.0,29.2),umidade.relativa=[49.3, 74.7)	$\Rightarrow$	precipitacao=Não Chuva	0.213	0.857	1.275	2.294	3.817
tempmaxima=[29.2,39.4],estacao=Outono	$\Rightarrow$	precipitacao=Não Chuva	0.062	0.856	1.272	2.270	3.081
umidade.relativa=[49.3, 74.7),velocidade.do.vento=[2.78,5.56)	$\Rightarrow$	precipitacao=Não Chuva	0.048	0.846	1.258	2.126	2.802
velocidade.do.vento=[2.78,5.56),estacao=Inverno	$\Rightarrow$	precipitacao=Não Chuva	0.025	0.831	1.236	1.940	2.450
tempmaxima=[29.2,39.4],umidade.relativa=[49.3, 74.7)	$\Rightarrow$	precipitacao=Não Chuva	0.235	0.819	1.218	1.808	2.845
umidade.relativa=[49.3, 74.7),tempminima=[17.53,26.40]	$\Rightarrow$	precipitacao=Não Chuva	0.232	0.800	1.190	1.639	2.450
tempmaxima=[29.2,39.4],estacao=Primavera	$\Rightarrow$	precipitacao=Não Chuva	0.112	0.794	1.180	1.587	2.046
tempmaxima=[19.0,29.2),umidade.relativa=[74.7,100.0],estacao=Verão	$\Rightarrow$	precipitacao=Chuva	0.046	0.791	2.417	3.225	8.908
tempmaxima=[8.8,19.0),tempminima=[8.67,17.53)	$\Rightarrow$	precipitacao=Chuva	0.014	0.786	2.402	3.149	7.868
velocidade.do.vento=[0.00,2.78),tempminima=[-0.20, 8.67)	$\Rightarrow$	precipitacao=Não Chuva	0.019	0.784	1.166	1.515	1.789
velocidade.do.vento=[0.00,2.78),estacao=Inverno	$\Rightarrow$	precipitacao=Não Chuva	0.167	0.779	1.158	1.481	1.951
tempmaxima=[19.0,29.2),estacao=Inverno	$\Rightarrow$	precipitacao=Não Chuva	0.128	0.770	1.145	1.422	1.776
tempmaxima=[19.0,29.2),umidade.relativa=[74.7,100.0],estacao=Primavera	$\Rightarrow$	precipitacao=Chuva	0.043	0.766	2.339	2.873	7.580
tempmaxima=[19.0,29.2),umidade.relativa=[74.7,100.0],tempminima=[17.53,26.40]	$\Rightarrow$	precipitacao=Chuva	0.092	0.748	2.285	2.670	8.102
tempmaxima=[19.0,29.2),tempminima=[17.53,26.40],estacao=Verão	$\Rightarrow$	precipitacao=Chuva	0.045	0.742	2.267	2.608	6.700
tempmaxima=[8.8,19.0),umidade.relativa=[74.7,100.0]	$\Rightarrow$	precipitacao=Chuva	0.016	0.740	2.260	2.586	6.091
umidade.relativa=[74.7,100.0],tempminima=[8.67,17.53),estacao=Primavera	$\Rightarrow$	precipitacao=Chuva	0.016	0.733	2.240	2.522	5.884
tempmaxima=[19.0,29.2),velocidade.do.vento=[0.00,2.78),estacao=Verão	$\Rightarrow$	precipitacao=Chuva	0.046	0.708	2.162	2.302	5.633

Aplicação I

Conclusões

- Sem redundância foram obtidas 32 regras.
- As regras com maiores confiança tem como consequência não chuva.
- O lift em todas as regras não foi alto, sendo o maior valor 2.42.
- Outras medidas (conviction e OR) indicam que a maioria dos eventos não são independentes,

Classificação

Definições

- Técnicas estatísticas multivariadas preocupadas com alocação de novos objetos (observações) em grupos previamente definidos (JOHNSON e WICHERN, 2007).
- Métodos utilizados:
  1. Classification Based on Associations (CBA).
  2. K-Nearest Neighbour (KNN).
  3. Linear Discriminat Analysis (LDA).
  4. Naive-Bayes (NB).

Classification Based on Associations (CBA)

Definição e Implementação

- Liu et al. (1998) propõe um algoritmo de classificação, baseado nas regras de associações obtidas.
- Variáveis preditoras contínuas devem ser discretizadas.
- Biblioteca `arulesCBA` do R tem implementada o algoritmo, há duas opções:
  - `CBA`: encontra a regra de associação e faz classificação.
  - `CBA_ruleset`: usa um conjunto de regras de associação pré-definidos para classificar.

K-Nearest Neighbour (KNN)

Conceito

Dado um novo vetor de observações x, para classificá-lo o método KNN tradicionalmente realiza as seguintes atividades:

1. Utilizando alguma medida de similaridade a distância entre a nova observação x e cada uma das observações do conjunto de dados é calculada.
2. As k observações mais próximas, isto é, mais similares a x são selecionadas.
3. A nova observação x é classificada na categoria mais frequente dos k vizinhos mais próximos.

K-Nearest Neighbour (KNN)

Definições e Implementação

- A distância euclidiana foi utilizada.
- O valor de k foi determinado avaliando a performance do classificador para $k = 2, 3, \ldots 12$.
- Implementação no R: biblioteca `classs` função `knn`.

Linear Discriminat Analysis (LDA)

Explicação

- Y: variável qualitativa com K classes, em que $K \geq 2$
- π_k: probabilidade de uma dada observação x estar associada a k-ésima categoria da variável resposta Y.
- $f_k(x) = \Pr[X = x \mid Y = k]$: função densidade de probabilidade de X para uma observação que vem da classe k.
- Então, o Teorema de Bayes afirma que:

$$p_k(x) = \Pr[Y = k \mid X = x] = \frac{\pi_k \, f_k(x)}{\sum_{i=1}^{K} \pi_i \, f_i(x)}. \qquad (1)$$

Linear Discriminant Analysis (LDA)

Explicação

- Assumiremos que as observações $X = (X_1, \ldots, X_p)$ da k-ésima classe seguem uma distribuição Normal multivariada $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$.
- Assim, o classificador LDA atribui uma observação $X = x$ à classe a qual

$$\delta_k(x) = x^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log \pi_k \qquad (2)$$

  é maior.
- Implementação no R: biblioteca `MASS` função `lda`.

Naive-Bayes (NB)

Explicação

- Mesma ideia do LDA, no entanto menos restritivo.
- O NB não faz restrição da distribuição de $f_k(x)$ e considera que as covariáveis (atributos) são condicionalmente independentes dado os valores das classes $Y = k$.
- Considerou-se uma forma não paramétrica via função de densidade Kernel para estimar $f_k(x)$
- Implementação no R: biblioteca `klaR` função `NaiveBayes`.

Avaliando a Performance

Medidas

1. Acurácia: proporção de observações (objetos) classificados corretamente.
2. Kappa: ajusta a acurácia, considerando a possibilidade de uma correta classificação ter sido por mero acaso.
3. Sensibilidade: proporção de registros “positivos” classificados corretamente.
4. Especificidade: proporção de registros “negativos” classificados corretamente.
5. Curva ROC: representação gráfica entre a sensibilidade e (1 - especificidade).

- Implementação no R: bibliotecas `caret` e `ROCR`.

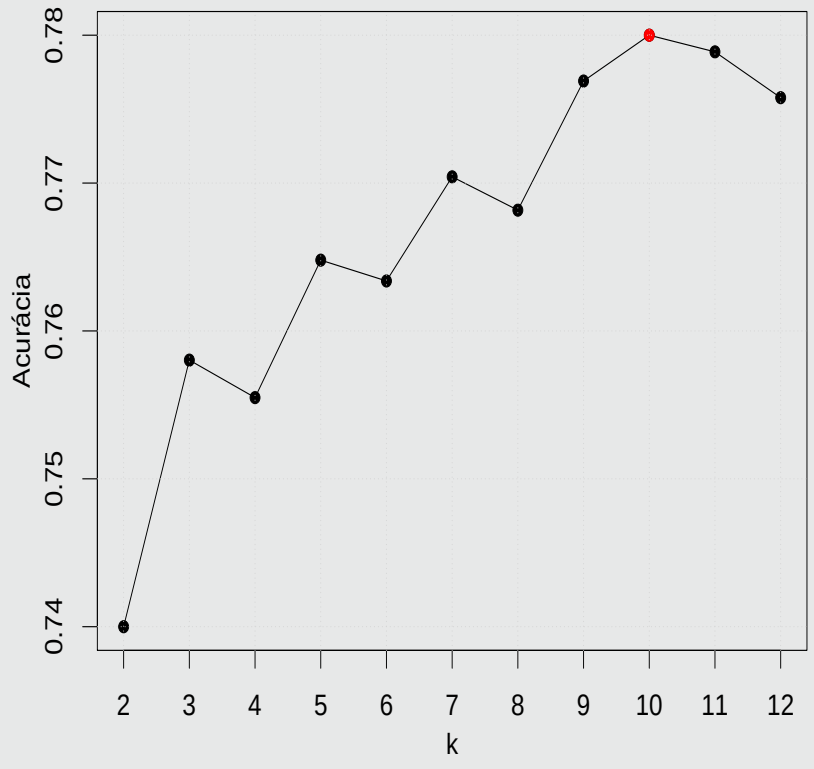
Aplicação I

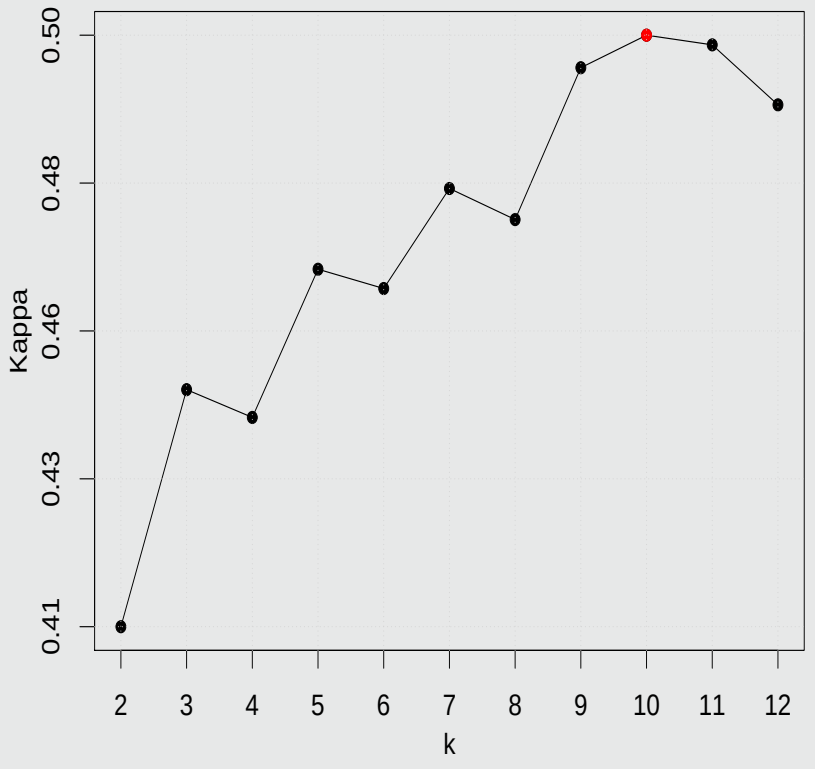
Problema

- Dada as informações sobre as variáveis climatológicas predizer se irá chover ou não.
- Métodos utilizados: CBA, KNN, LDA e NB.
- Performance avaliada via: acurácia, Kappa e sensibilidade.
- Dados foram divididos em 70% treinamento e 30% teste;.
- $B = 1000$ amostras Bootstrap foram utilizadas para avaliar a precisão das medidas.

Aplicação I

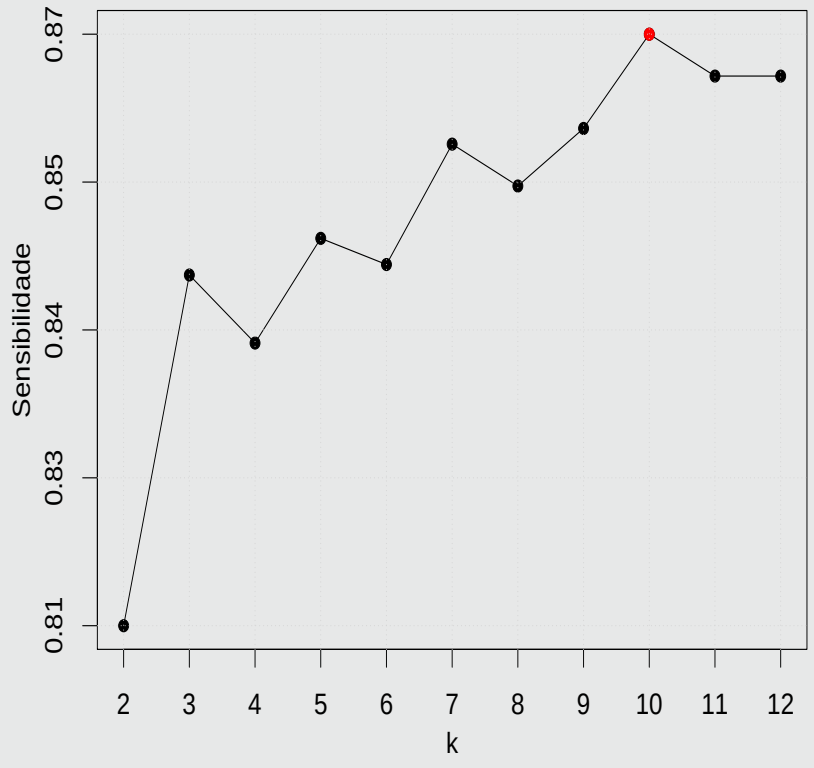
Determinando valor de k no KNN

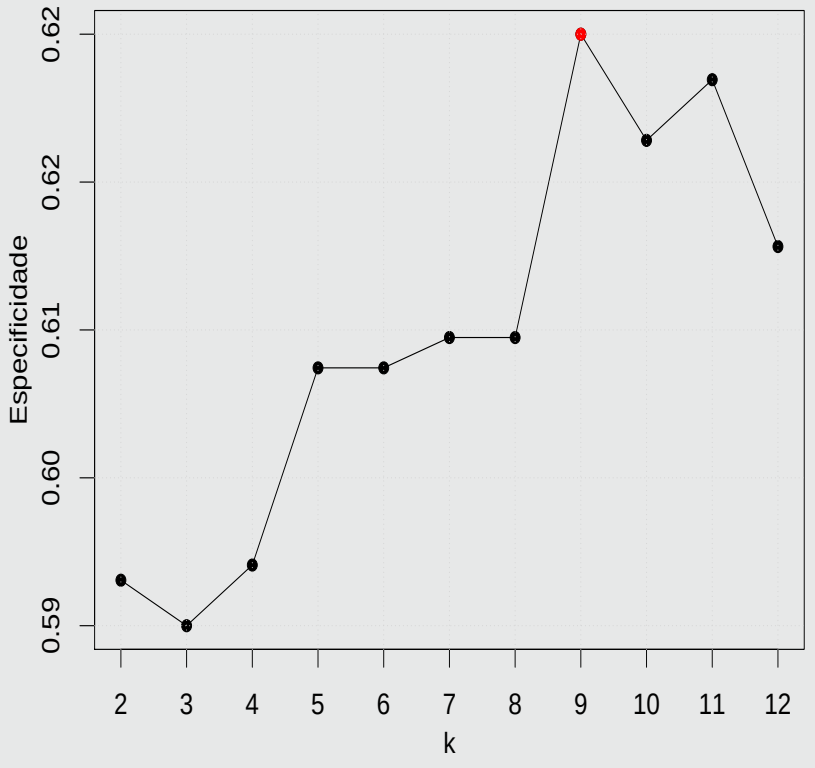




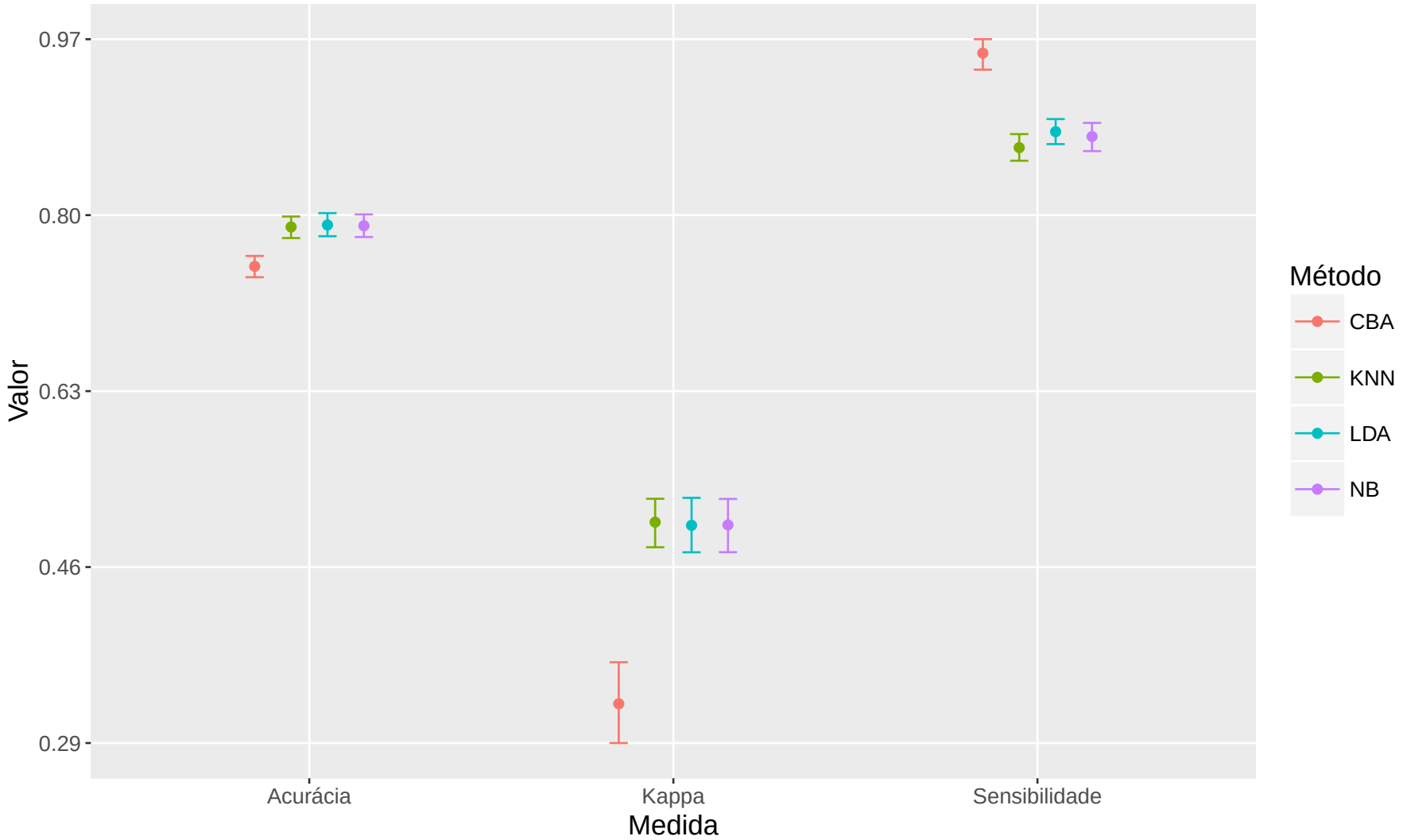
Aplicação I

Determinando valor de k no KNN





Aplicação I


0.97
0.80
0.63
0.46
0.29
Valor
Acurácia
Kappa
Sensibilidade
Medida
Método
CBA
KNN
LDA
NB

Aplicação I

Performance dos métodos de classificação

Método	Acurácia			Kappa			Sensibilidade			Rank
	LI	Média	LS	LI	Média	LS	LI	Média	LS	
CBA	0.7374	0.7479^{4}	0.7578	0.2903	0.3280^{4}	0.3678	0.9367	0.9525^{1}	0.9659	9
KNN	0.7750	0.7857^{3}	0.7957	0.4782	0.5023^{1}	0.5247	0.8493	0.8618^{4}	0.8748	8
LDA	0.7768	0.7876^{1}	0.7990	0.4734	0.4993^{3}	0.5256	0.8652	0.8772^{2}	0.8893	6
NB	0.7760	0.7868^{2}	0.7977	0.4735	0.4996^{2}	0.5246	0.8585	0.8725^{3}	0.8856	7

LI(S) limite inferior (superior) do intervalo de 95% de confiança Bootstrap.

Aplicação I

Conclusões

- LDA apresentou melhor desempenho;
- CBA apresentou pior desempenho em termo de Acurácia, no entanto mostro-se melhor em termos de Especificidade.
- Os métodos LDA, KNN e NB apresentaram resultados muito semelhantes.

Aplicação II

Problema

- Motivação: trabalho de Sipko (2015).
- Objetivo: prever o resultado de partidas de tenis.
- Dados: 31.354 partidas da ATP de 2004 a 2015
  `https://github.com/okh1/tennis-prediction`
- Método utilizado: Naive-Bayes.
- Performance avaliada via: acurácia, Kappa, sensibilidade, especificidade e curva ROC.

Aplicação II

Dados

Series	Court	Surface	Round	Best of	Rank	B365	CB	EX	IW	PS	Result
International	Outdoor	Hard	1st Round	3	29	NA	1.40	1.48	1.45	1.47	1
International	Outdoor	Hard	1st Round	3	79	NA	2.85	2.53	2.20	2.90	0
International	Outdoor	Hard	1st Round	3	33	1.16	1.22	1.20	1.20	1.24	1
International	Outdoor	Hard	1st Round	3	64	4.50	4.10	4.45	3.30	4.55	0
International	Outdoor	Hard	1st Round	3	54	2.00	2.15	NA	2.00	2.17	1
International	Outdoor	Hard	1st Round	3	74	1.72	1.67	NA	1.55	1.75	0
International	Outdoor	Hard	1st Round	3	66	1.83	1.70	NA	NA	1.73	1
International	Outdoor	Hard	1st Round	3	75	1.83	2.10	NA	NA	2.21	0
International	Outdoor	Hard	1st Round	3	36	1.40	1.40	1.50	1.35	1.46	1
International	Outdoor	Hard	1st Round	3	82	2.75	2.85	2.45	2.50	2.93	0
.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	
Masters Cup	Indoor	Hard	Round Robin	3	5	1.20	NA	1.20	NA	1.24	1
Masters Cup	Indoor	Hard	Round Robin	3	7	4.50	NA	4.30	NA	4.57	0
Masters Cup	Indoor	Hard	Round Robin	3	4	2.30	NA	2.20	NA	2.34	1
Masters Cup	Indoor	Hard	Round Robin	3	2	1.61	NA	1.65	NA	1.68	0
Masters Cup	Indoor	Hard	Semifinals	3	1	1.20	NA	1.18	NA	1.22	1
Masters Cup	Indoor	Hard	Semifinals	3	5	4.50	NA	4.30	NA	4.92	0
Masters Cup	Indoor	Hard	Semifinals	3	3	1.30	NA	1.33	NA	1.30	1
Masters Cup	Indoor	Hard	Semifinals	3	4	3.50	NA	3.20	NA	4.01	0
Masters Cup	Indoor	Hard	The Final	3	1	1.44	NA	1.42	NA	1.40	1
Masters Cup	Indoor	Hard	The Final	3	3	2.75	NA	2.80	NA	3.27	0

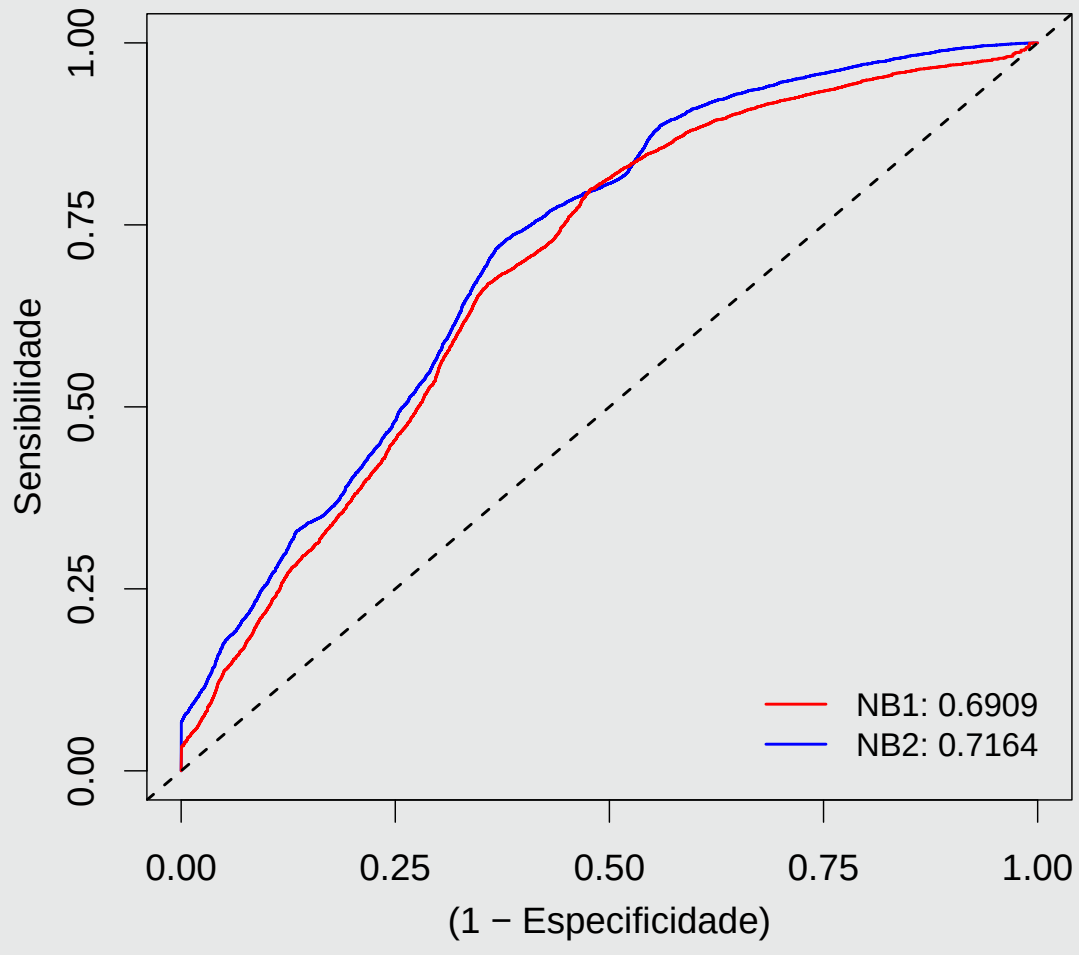
Aplicação II

Método

- Para predizer o resultado de uma partida de tenis duas abordagens do Naive-Bayes foram adotadas:
  1. NB1: desconsidera os registros faltantes.
  2. NB2: desconsidera apenas os atributos faltantes.

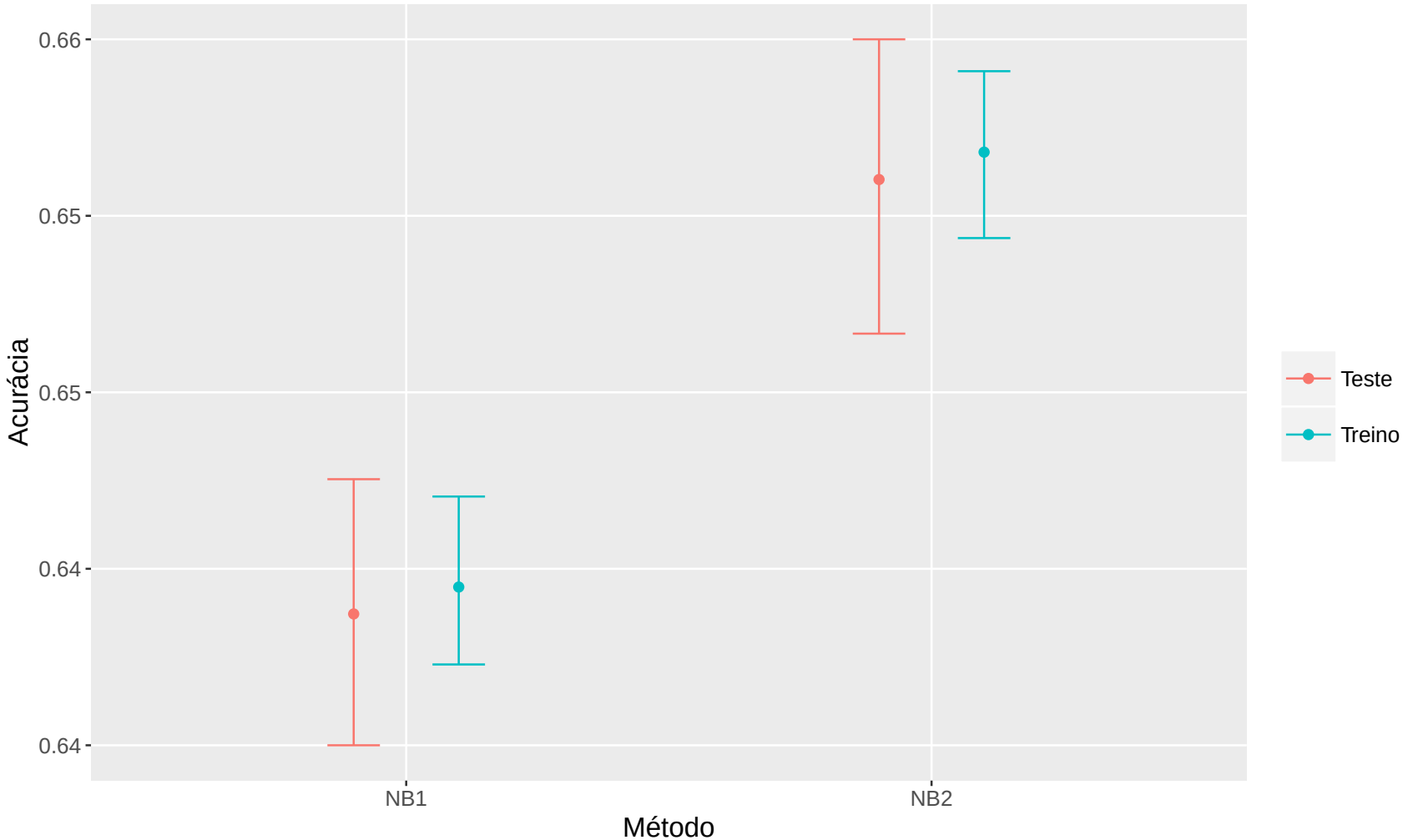
Aplicação II

Curva ROC



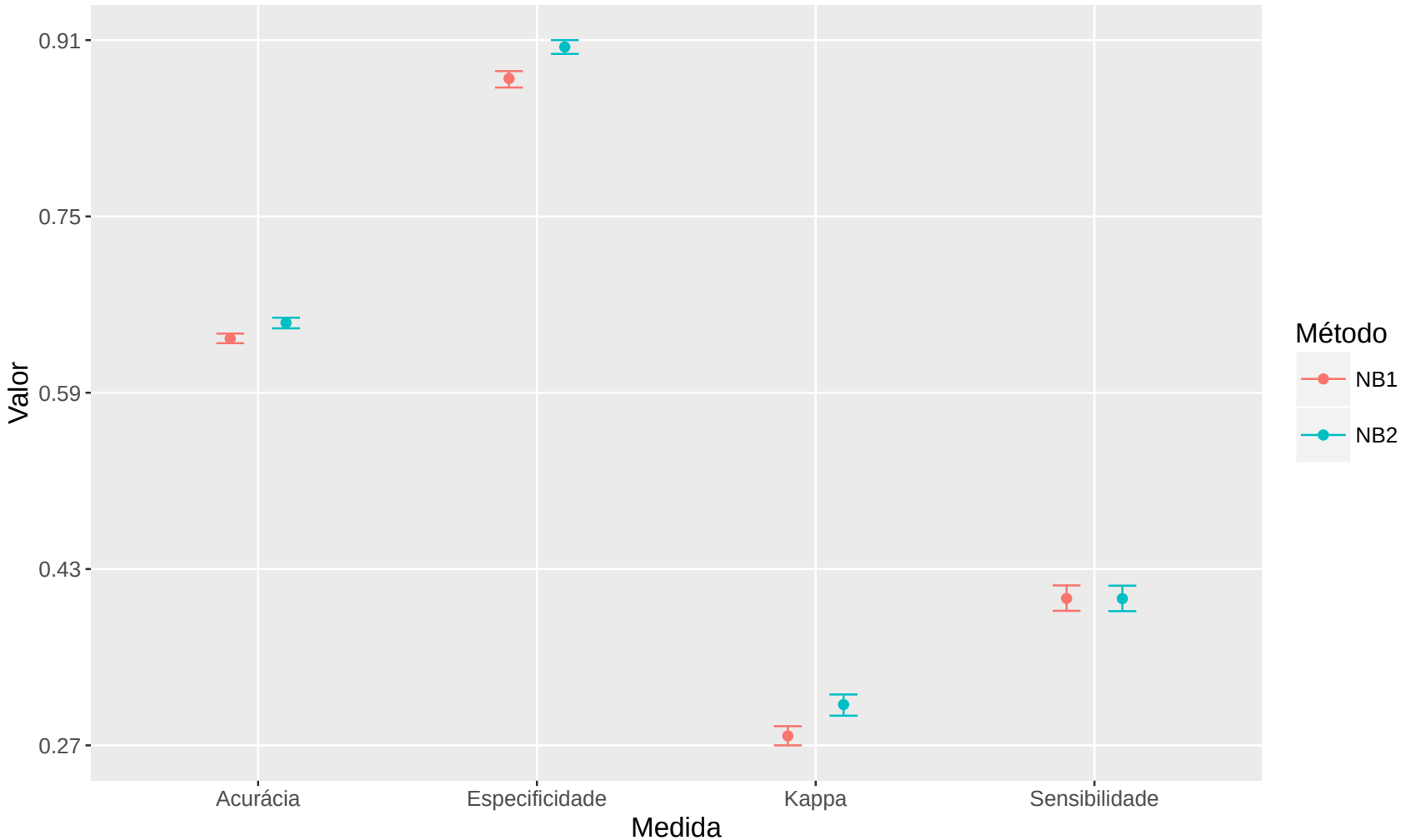
Aplicação II

- *Over-fit*.



Aplicação II

- Performance.



Aplicação II

Performance – Média das Medidas

Método	Acurácia	Kappa	Sensibilidade	Especificidade
NB1	0.6407	0.2814	0.4057	0.8757
NB2	0.6548	0.3097	0.4055	0.9042

Referências

Referências

[1] Agrawal, R., Srikant, R., 1994. **Fast algorithms for mining association rules in large databases.** In: Proceedings of the 20th International Conference on Very Large Data Bases. VLDB '94. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 487–499.

[2] James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. **An Introduction to Statistical Learning with Applications in R.** Springer.

[3] Johnson, R. A., Wichern, D. W., 2007. **Applied Multivariate Statistical Analysis**, Sixth Edition. Prentice Hall.

Referências

Referências

[4] Liu, B., Hsu, W., Ma, Y., 1998. **Integrating classication and association rule mining.** In: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining. KDD'98. AAAI Press, pp. 80–86.

[5] Sipko, M., 2015. **Machine Learning for the Prediction of Professional Tennis Matches.** Technical report, Imperial College London, London.