

Aplicação de Regras de Associação e Métodos de Classificação em Variáveis Climatológicas

André Felipe Berdusco Menezes

Departamento de Estatística, Universidade Estadual de Maringá

Maringá, PR, Brazil

1 Materiais e Métodos

1.1 Conjunto de Dados

Para realizar este trabalho, foi utilizado os dados relacionados as variáveis climatológicas da estação de Maringá durante os anos de 1961 a 2016. Os dados foram obtidos a partir do Instituto Nacional de Meteorologia (INMET) <http://www.inmet.gov.br/> e consistem de 16.062 registros. Foram consideradas seis variáveis: estação do ano, precipitação (mm), umidade relativa (%), velocidade do vento (mps), temperatura máxima (°C) e temperatura mínima (°C). Tanto para as regras de associação quanto para a classificação a variável de interesse foi se ocorreu chuva ou não, assim utilizando a variável precipitação definimos uma nova variável dicotômica, indicando se chuva para níveis de precipitações maiores que zero e não chuva caso contrário. Algumas medidas descritivas das variáveis não categorizadas estão expostas na Tabela 1.

Tabela 1: Medidas descritivas das variáveis.

Variável	Faltantes	Mín.	Média	Mediana	Máx.	DP	Skw.	Kurt.
Precipitação (mm)	61	0.00	4.60	0.00	170.30	11.94	4.15	23.34
Umidade Relativa (%)	71	24.00	69.21	69.00	100.00	14.39	-0.09	-0.61
Velocidade do Vento (mps)	0	0.00	1.41	1.00	8.33	1.02	1.74	4.06
Temperatura Máxima (°C)	137	8.80	28.10	28.70	39.40	4.14	-0.71	0.52
Temperatura Mínima (°C)	46	-0.20	17.48	18.20	26.40	3.82	-0.98	1.10

Uma vez que os dados estão na forma de série temporal utilizamos interpolação linear para substituir os valores faltantes nas variáveis. A título de ilustração, exibimos na Figura 1 o comportamento antes e após a interpolação da variável umidade relativa em função do tempo.

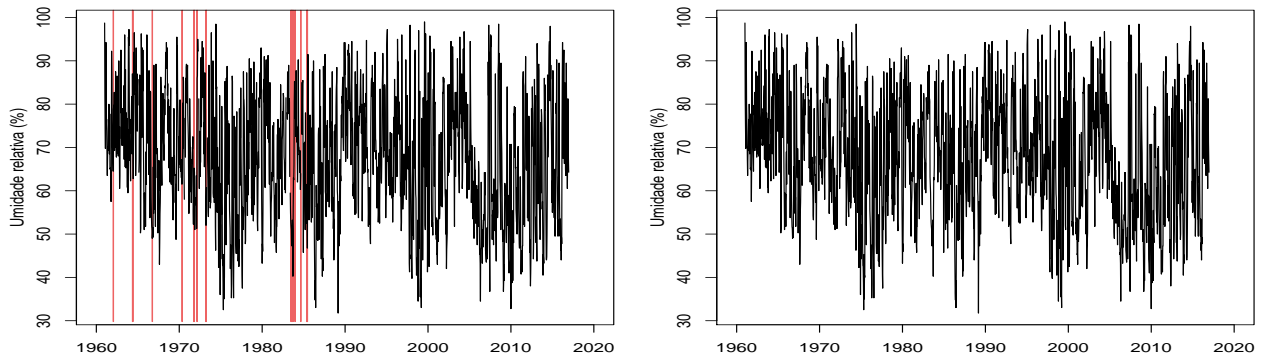


Figura 1: Gráficos da série temporal da umidade relativa antes e após a interpolação.

É importante mencionar que para realizar as regras de associação foi necessário discretizar as variáveis contínuas. Na Tabela 2 exibimos as frequências absolutas e percentuais para as variáveis categorizadas. Um fato interessante é que durante 1963 e 2016 ocorreram mais dias sem chuva do que com na cidade de Maringá.

Tabela 2: 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)

1.2 Regras de Associação

Em virtude do advento da informação grandes bases de dados estão disponíveis hoje em dia, e a mineração de dados (*data mining*) vem sendo útil para analisar e extrair o máximo de informações destes conjuntos de dados. Dentre as diversas técnicas presentes na mineração de dados as regras de associação, ou em inglês *association rules*, é bastante utilizada para descobrir relações interessantes entre variáveis em grandes bancos de dados. Seguramente, a análise da cesta de mercado (*Market Basket Analysis*) é o principal exemplo de aplicação, onde as regras de associação são empregadas para fazer recomendações de produtos, identificando produtos que são frequentemente comprados juntos.

Agrawal et al. (1993) introduziram o conceito de regras de associação da seguinte forma. Seja $I = \{i_1, i_2, \dots, i_m\}$ um conjunto m atributos (características) dicotômicos denominado **items**. Seja $D = \{t_1, t_2, \dots, t_n\}$ um conjunto de transações (observações) denominado **database** (ou banco de dados). Cada transação em D contém um subconjunto de **items** em I , isto é, $t \subset I$. Uma regra é definida como uma implicação da forma $X \Rightarrow Y$, em que $X, Y \subset I$ e $X \cap Y = \emptyset$. O conjunto de itens (abreviado **itemsets**) X e Y são chamados de antecedente (left-hand-side ou LHS) e conseqüente (right-hand-side ou RHS) da regra. As três medidas mais importantes utilizadas para seleção de regras de associação são: *support*, *confidence* e *lift*. Sob o ponto de vista probabilístico, essas medidas, são estimativas empíricas das probabilidades associada aos eventos (regras) $A \Rightarrow B$. Elas são definidas como:

$$\text{support}(A \Rightarrow B) = P(A \cap B), \quad (1)$$

$$\text{confidence}(A \Rightarrow B) = \frac{P(A \cap B)}{P(A)}, \quad (2)$$

$$\text{lift}(A \Rightarrow B) = \frac{P(A \cap B)}{P(A) P(B)}. \quad (3)$$

Usualmente, os conjuntos de dados são extramente grandes em termos da quantidade de observações (transações) e do número de características (**items**) (Lantz, 2013). Assim Agrawal and Srikant (1994) desenvolveram um algoritmo denominado APIORI, que permite selecionar as regras tais que satisfazem um limite inferior para o suporte e confiança. Este algoritmo esta implementado na biblioteca **arules** (Hahsler et al., 2005) do software R.

O suporte e a confiança fornecem estimativas empíricas para as probabilidades do eventos (regras). Por outro lado, valores de lift próximos de um indicam que a regra $A \Rightarrow B$ são independentes. Na biblioteca **arules** estão disponíveis 42 medidas, além do suporte, confiança e lift. Assim, neste trabalho também foram utilizadas duas outras medidas, afim auxiliar o lift em detectar a existência de independência entre as regras. Elas são definidas como:

$$\text{conviction}(A \Rightarrow B) = \frac{P(A) P(\bar{B})}{P(A \cap \bar{B})} \quad (4)$$

$$\text{OR}(A \Rightarrow B) = \frac{P(A \cap B) P(\bar{A} \cap \bar{B})}{P(A \cap \bar{B}) P(\bar{A} \cap B)} \quad (5)$$

A primeira medida foi proposta por Brin et al. (1997) e assim como a razão de chances, ou Odds Ratio (OR), indica independência quando for igual a um.

1.3 Métodos de Classificação

Em sua essência classificação são abordagens para prever variáveis respostas qualitativas dado um conjunto de variáveis explicativas. Como aponta James et al. (2013) prever uma resposta qualitativa para uma dada observação pode ser referido como classificando, uma vez que envolve a atribuição da observação a uma categoria ou classe. Por outro lado, muitas vezes os métodos utilizados para classificação primeiro predizem a probabilidade de cada uma das categorias de uma variável qualitativa, e assim realizam a classificação.

Johnson and Wichern (2007) define classificação como técnicas estatísticas multivariadas preocupadas com alocação de novos objetos (observações) em grupos previamente definidos. Assim, os autores estabelecem que o objetivo imediato da classificação é: classificar um objeto (observação) em duas ou mais classes rotuladas. A ênfase está em derivar uma regra que pode ser usada para atribuir de forma otimizada novos objetos às classes rotuladas.

Nesta seção iremos discutir brevemente os métodos de classificação utilizados neste trabalho. Além disso, será apresentada as medidas empregadas para avaliar a performance dos métodos de classificação.

1.3.1 Classificação baseada em associações

Proposto por Liu et al. (1998) a classificação baseada em regras de associação consiste em determinar (i) uma regra geradora, isto é, um conjunto de regras de associações, baseada no algoritmo Apriori; (ii) e um método para construir um classificador. No software R a biblioteca `arulesCBA` (Johnson, 2017) apresenta o algoritmo CBA (Classification Based on Associations) para construir um classificador considerando um conjunto de dados discretos.

1.3.2 Classificação baseada nos k vizinhos mais próximos

Um método não paramétrico bastante utilizado para o problema de classificação é o classificador baseado nos k vizinhos mais próximos, em inglês k -nearest neighbour (KNN) (Cover and Hart, 1967).

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

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

Ao realizar os passos descritos, surgem dois possíveis questionamentos. (i) determinar uma função de distância (medida de similaridade); (ii) determinar o número de vizinhos, isto é, valor de k . Por questões de conveniência, neste trabalho foi a distância euclidiana foi utilizada. Para determinar o número de vizinhos, a performance do KNN foi avaliada em termos de acurácia e outras medidas para os seguintes valores $k = 2, 3, \dots, 10$. Ressalta-se que o KNN está implementado no software R na biblioteca `class` (Venables and Ripley, 2002). No software SAS o KNN está disponível na PROC DISCRIM (SAS Institute Inc., 2011).

1.3.3 Análise discriminante linear

Suponha que desejamos classificar uma observação em uma das classes K , em que $K \geq 2$. Em outras palavras, a variável resposta qualitativa Y pode assumir K possíveis distintos e desordenados valores. Considere que π_k representa a probabilidade a priori de uma observação escolhida aleatoriamente venha da k -ésima classe, ou seja a probabilidade de que uma dada observação está associada com a k -ésima categoria da variável resposta Y . Considere que $f_k(x) = \Pr[X = x | Y = k]$ denota a função densidade de probabilidade de X para uma observação que vem da classe k . Então, o Teorema de Bayes afirma que:

$$p_k(x) = \Pr[Y = k | X = x] = \frac{\pi_k f_k(x)}{\sum_{i=1}^K \pi_i f_i(x)}. \quad (6)$$

Em geral, estimar π_k é fácil se tivermos uma amostra aleatória da população Y basta calcular a proporção de observações que pertencem a k -ésima classe. No entanto, estimar $f_k(x)$ tende a ser mais desafiado, a menos que assumimos algumas formas mais simples para essas densidades. Portanto, queremos obter uma estimativa para $f_k(x)$ para utilizar em (6) afim de estimar $p_k(x)$. Em seguida, classificaremos uma observação para uma classe tal que $p_k(x)$ é maior.

Na análise discriminante linear (LDA), proposta por Fisher (1938), assumiremos que as observações $X = (X_1, \dots, X_p)$ da k -ésima classe seguem uma distribuição Normal multivariada $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, em que $\boldsymbol{\mu}_k$ é um vetor de médias específico da classe e $\boldsymbol{\Sigma}$ é uma matriz de covariância

comum para as K classes. Formalmente, a densidade Normal multivariada é definida como:

$$f(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right) \quad (7)$$

Substituindo (7) na função densidade para a classe k em (6) e realizando algumas manipulações algébricas, pode-se mostrar que o classificador LDA atribui uma observação $X = x$ à classe a qual

$$\delta_k(x) = x^\top \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^\top \Sigma^{-1} \mu_k + \log \pi_k \quad (8)$$

é maior. Note que $\delta_k(x)$ é uma função linear de x , este é o motivo do termo *linear* no nome deste método. Ressalta-se que o LDA está implementado no software R na biblioteca MASS (Venables and Ripley, 2002). No software SAS o LDA está disponível na PROC DISCRIM (SAS Institute Inc., 2011)

1.3.4 Naive-Bayes

Assim como o LDA o classificador Naive-Bayes (NB) utiliza o Teorema de Bayes, ou seja, uma nova observação $X = x$ é classificada a classe com maior probabilidade a posteriori $p_k(x)$. No entanto, ao contrário do LDA que assume distribuição Normal multivariada para $f_k(x) = \Pr[X = x | Y = k]$ 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 $Y = y$.

Existem uma infinidade de distribuições de probabilidade que podem ser assumidas para os atributos. Conforme discutido em Pérez et al. (2009) para estimar a função densidade, $f_k(x)$, as seguintes abordagens são geralmente consideradas:

1. Discretizar as variáveis contínuas e estimar a distribuição de probabilidade utilizando a distribuição multinomial;
2. Utilizar um modelo paramétrico, por exemplo a distribuição Normal;
3. Utilizar um modelo não paramétrico, por exemplo as funções de densidade Kernel.

Nesse sentido, o presente trabalho considera a forma não paramétrica via função de densidade Kernel. Ressalta-se que o NB considerando a função densidade Kernel para $f_k(x)$ está implementado na biblioteca klaR (Weihs et al., 2005) do software R. Já no SAS o NB está disponível na PROC DISCRIM (SAS Institute Inc., 2011).

2 Resultados e Discussões

2.1 Regras de Associação

A Figura 2 exibe as frequências absoluta e percentuais dos itemsets (eventos). Pode-se verificar que o item mais frequente é a velocidade do vento entre $[0.00, 2.78)$, enquanto que o item menos frequente é a velocidade do vento entre $[5.56, 8.33]$, ocorrendo em somente 97 dias.

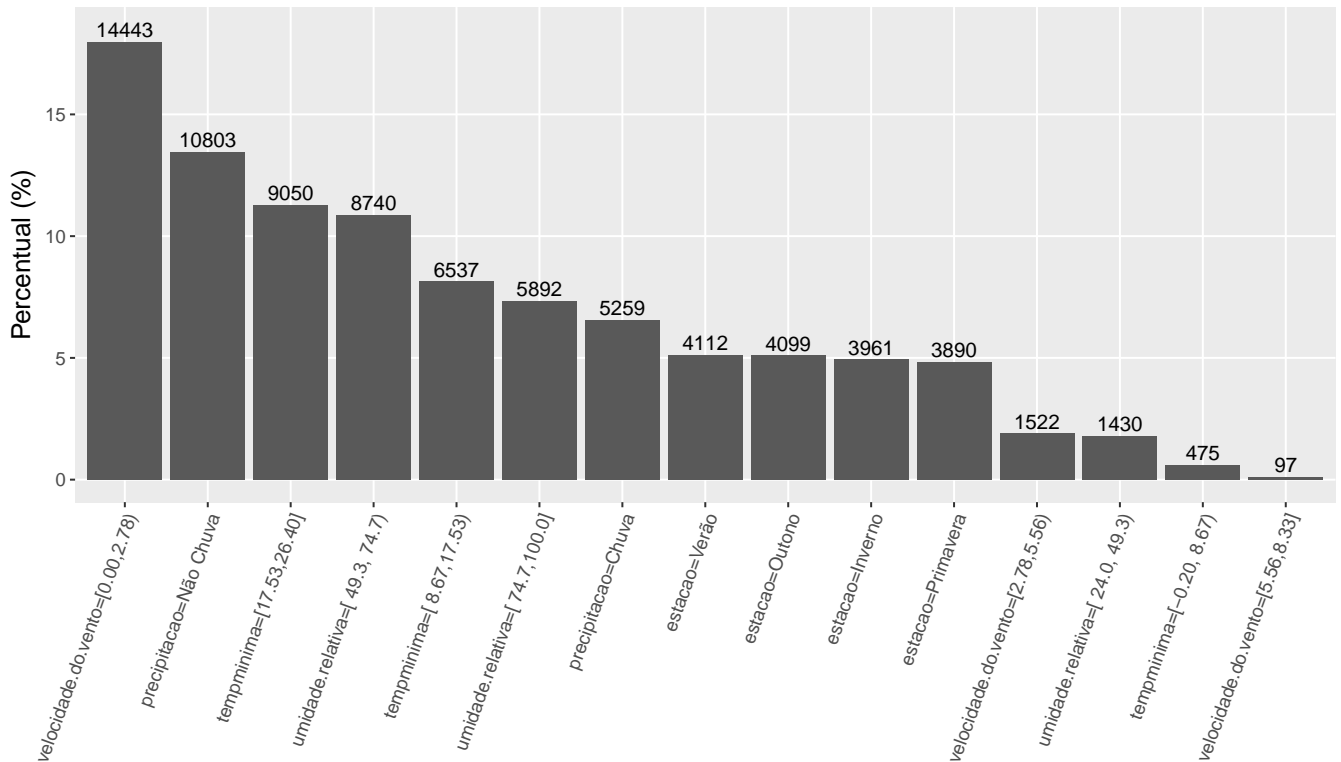


Figura 2: Frequências absoluta e percentuais dos itemsets.

Uma representação gráfica entre suporte, confiança e lift é apresentada na Figura 3. Nesta figura podemos notar que as regras com maiores confianças apresentam baixo suporte, entre 0.01 e 0.11, e lift próximo de 1.44. Observa-se também que as regras com maiores lifts, aproximadamente 2.42, apresentam baixa confiança e suporte.

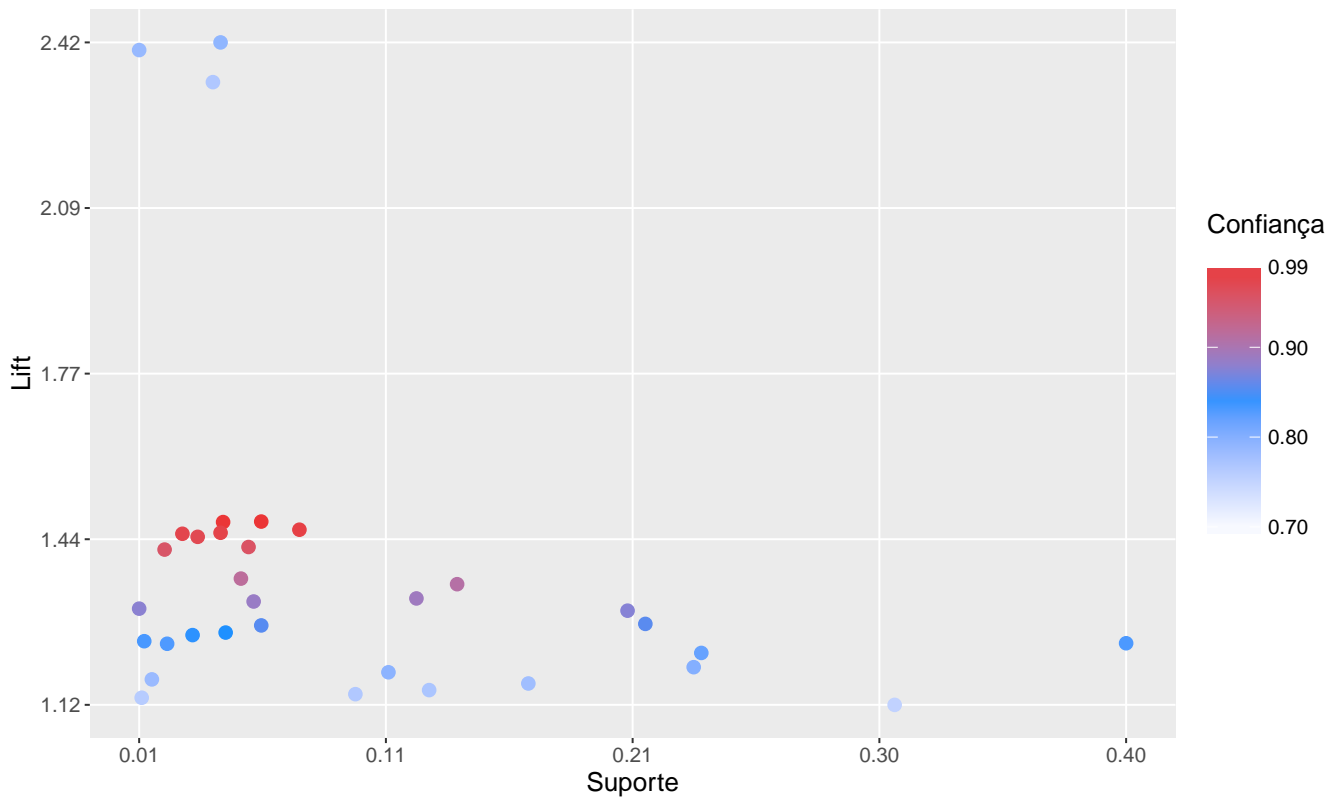


Figura 3: Relação entre suporte, confiança e lift das regras de associação encontradas.

Afim de obter notáveis regras de associação entre as variáveis foi estabelecido limites mínimos de suporte e confiança de 0.01 e 0.75, respectivamente. Os resultados das regras de associação, ordenados conforme a confiança, estão expostos na Tabela 3.

Diversas conclusões podem ser extraídas da Tabela 3, destacamos as seguintes:

1. Ao todo e sem redundância foram obtidas 33 regras.
2. As regras com maiores confiança tem como consequência chuva.
3. O lift em todas as regras não foi alto, sendo o maior valor 2.42. No entanto, as outras medidas (conviction e OR) indicam que a maioria dos eventos não são independentes,
4. A probabilidade de não chover dado que a umidade relativa esteve entre $[24.0, 49.3)$ e a temperatura máxima entre $[29.2, 39.4]$ é de 99,3%.
5. A probabilidade de não chover dado que a temperatura máxima esteve entre $[29.2, 39.4]$, e a estação do ano é o inverno é aproximadamente 95,9%.
6. A probabilidade de chover dado que estamos no verão, a temperatura máxima esteve entre $[19.0, 29.2)$ e a umidade relativa variou entre $[74.7, 100.0]$ é de 79,1%.
7. A probabilidade de não chover na primavera e inverno são muito similares quando a umidade relativa varia entre $[24.0, 49.3)$.

8. A probabilidade de chover dado que estamos no Outono e a velocidade do vento esteve entre $[2.78, 5.56)$ é de 75%.

Tabela 3: Regras de associação.

Regra ($A \Rightarrow B$)	SupORTE	Confiança	Lift	Conviction	OR
tempmaxima=[29.2,39.4],umidade.relativa=[24.0, 49.3)	0.062	0.993	1.476	47.195	76.705
umidade.relativa=[24.0, 49.3),tempminima=[17.53,26.40]	0.047	0.992	1.475	41.582	65.878
umidade.relativa=[24.0, 49.3),velocidade.do.vento=[0.00,2.78)	0.077	0.982	1.460	17.965	29.492
umidade.relativa=[24.0, 49.3),estacao=Inverno	0.046	0.978	1.454	14.695	22.873
umidade.relativa=[24.0, 49.3),estacao=Primavera	0.031	0.976	1.452	13.861	21.042
umidade.relativa=[24.0, 49.3),tempminima=[8.67,17.53)	0.037	0.972	1.446	11.826	18.037
tempmaxima=[29.2,39.4],estacao=Inverno	0.057	0.959	1.426	8.060	12.474
tempmaxima=[19.0,29.2],umidade.relativa=[24.0, 49.3)	0.024	0.956	1.421	7.367	10.818
tempminima=[17.53,26.40],estacao=Inverno	0.054	0.917	1.364	3.958	5.781
umidade.relativa=[49.3, 74.7),estacao=Outono	0.139	0.910	1.353	3.640	5.950
umidade.relativa=[49.3, 74.7),estacao=Inverno	0.123	0.891	1.325	3.017	4.673
tempmaxima=[29.2,39.4],tempminima=[8.67,17.53)	0.059	0.887	1.319	2.909	4.110
umidade.relativa=[49.3, 74.7),tempminima=[-0.20, 8.67)	0.014	0.878	1.305	2.681	3.553
tempmaxima=[19.0,29.2],tempminima=[8.67,17.53)	0.206	0.875	1.301	2.615	4.458
tempmaxima=[19.0,29.2],umidade.relativa=[49.3, 74.7)	0.213	0.857	1.275	2.294	3.817
tempmaxima=[29.2,39.4],estacao=Outono	0.062	0.856	1.272	2.270	3.081
umidade.relativa=[49.3, 74.7),velocidade.do.vento=[2.78,5.56)	0.048	0.846	1.258	2.126	2.802
tempmaxima=[29.2,39.4],velocidade.do.vento=[2.78,5.56)	0.035	0.843	1.253	2.084	2.700
umidade.relativa=[49.3, 74.7),tempminima=[-0.20, 8.67)	0.016	0.834	1.241	1.977	2.489
umidade.relativa=[49.3, 74.7),velocidade.do.vento=[0.00,2.78)	0.402	0.832	1.237	1.948	4.509
velocidade.do.vento=[2.78,5.56),estacao=Inverno	0.025	0.831	1.236	1.940	2.450
tempmaxima=[29.2,39.4],umidade.relativa=[49.3, 74.7)	0.235	0.819	1.218	1.808	2.845
umidade.relativa=[49.3, 74.7),tempminima=[17.53,26.40]	0.232	0.800	1.190	1.639	2.450
tempmaxima=[29.2,39.4],estacao=Primavera	0.112	0.794	1.180	1.587	2.046
tempmaxima=[19.0,29.2],umidade.relativa=[74.7,100.0],estacao= Verão	0.046	0.791	2.417	3.225	8.908
tempmaxima=[8.8,19.0),tempminima=[8.67,17.53)	0.014	0.786	2.402	3.149	7.868
velocidade.do.vento=[0.00,2.78),tempminima=[-0.20, 8.67)	0.019	0.784	1.166	1.515	1.789
velocidade.do.vento=[0.00,2.78),estacao=Inverno	0.167	0.779	1.158	1.481	1.951
tempmaxima=[19.0,29.2],estacao=Inverno	0.128	0.770	1.145	1.422	1.776
tempmaxima=[19.0,29.2],umidade.relativa=[74.7,100.0],estacao=Primavera	0.043	0.766	2.339	2.873	7.580
umidade.relativa=[49.3, 74.7),estacao=Primavera	0.099	0.765	1.137	1.392	1.683
velocidade.do.vento=[2.78,5.56),estacao=Outono	0.015	0.760	1.130	1.366	1.556
tempmaxima=[29.2,39.4],velocidade.do.vento=[0.00,2.78)	0.311	0.751	1.116	1.314	1.869

2.2 Classificação

Esta seção é dedicada a aplicação dos métodos CBA, KNN, LDA e Naive-Bayes para predição de chuva em um determinado dia na cidade de Maringá. Para comparar a performance entre os classificadores, três medidas foram utilizados, a acurácia, coeficiente Kappa e sensibilidade (Lantz, 2013). Além disso, o procedimento Bootstrap com $B = 1000$ repetições foi empregado para estimar a precisão dessas medidas. As análises foram implementadas no software R e a biblioteca `caret` (Kuhn, 2017) foi utilizada extensivamente para estimar as medidas.

No classificador KNN é necessário determinar um valor de k . Para não tomar uma decisão arbitrária, a performance, em termo de acurácia, sensibilidade e especificidade, do KNN foi avaliada considerando $k = 2, 3, \dots, 12$. A Figura 4 mostra o comportamento dessa medidas para os diferente valores de k . Observa-se que os maiores valores são atingidos para $k = 10$, exceto para a especificidade em que possui maior valor para $k = 9$. Assim sendo, será considerado $k = 10$.

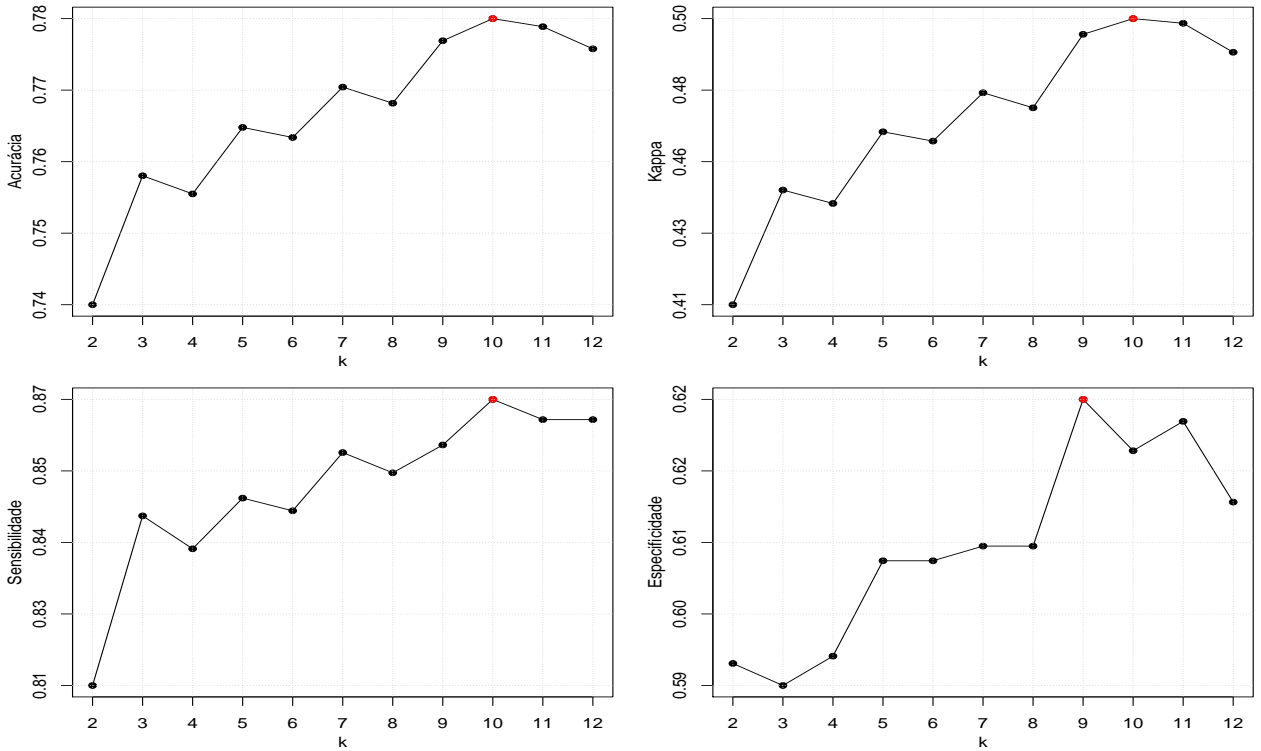


Figura 4: Determinando o valor de k no classificador KNN.

Para avaliar a performance dos classificadores, utilizou-se o procedimento Bootstrap com $B = 1000$ e considerando 70% dos dados como treino. A média das métricas utilizadas bem como os intervalos de confiança Bootstrap estão apresentados na Tabela 4 e Figura 5. Verifica-se que o método LDA obteve os melhores resultados de acurácia e Kappa. Por outro lado, o classificador CBA obteve melhor sensibilidade, isto é, ele classifica melhor os dias com chuva dos que demais classificadores.

Tabela 4: 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.

Na Figura 5 temos uma representação gráfica do desempenho de cada classificador. Embora, o LDA, em geral, apresentou melhor desempenho em termos de acurácia fica evidente que os métodos KNN e NB foram muito similar a ele.

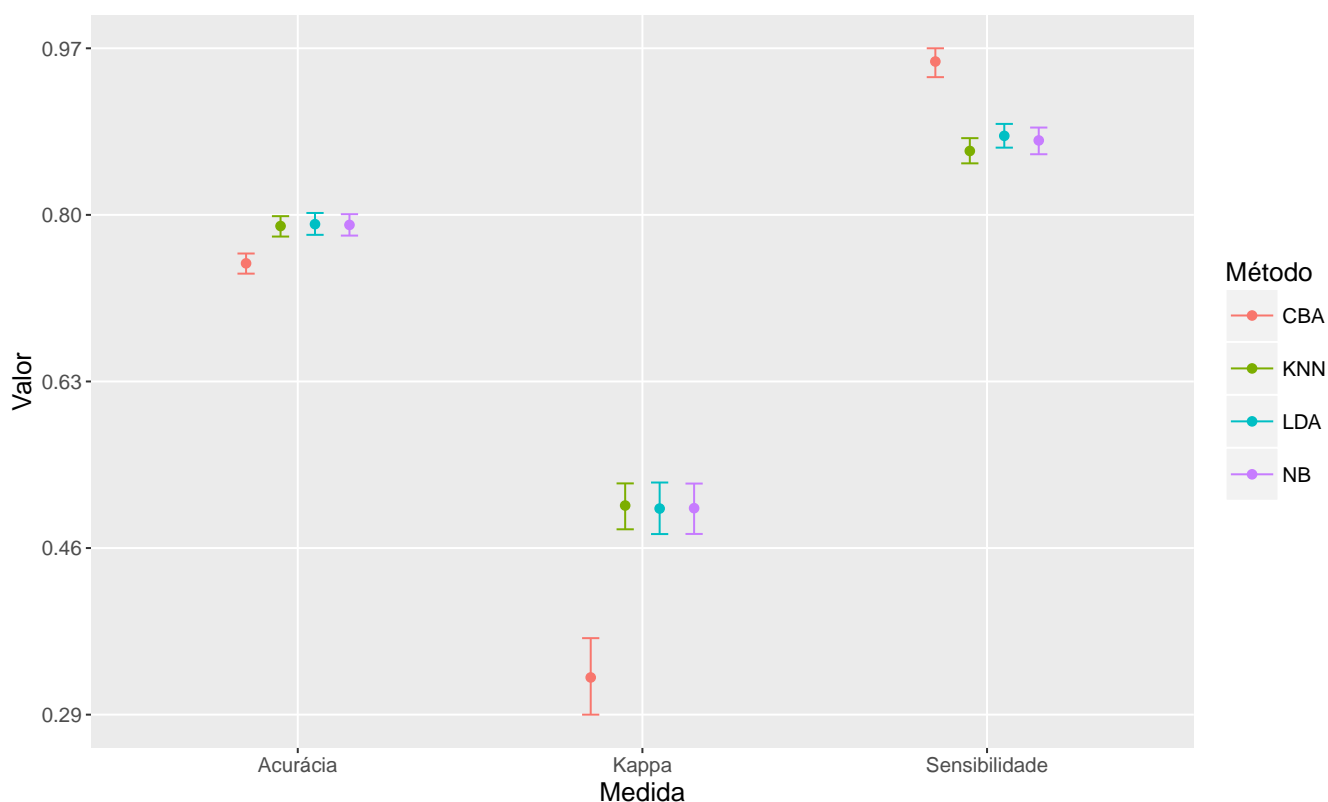


Figura 5: Performance dos métodos de classificação.

Pode-se concluir que os classificadores KNN, LDA e NB exibiram desempenho muito similares, com acurácia de aproximadamente de 78%. Isto significa que a cada 100 dias, é esperado que o modelo acerte 78 desfechos, isto é, se vai chover ou não.

Referências

- Agrawal, R., Imieliński, T., Swami, A., 1993. Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. SIGMOD '93. ACM, New York, NY, USA, pp. 207–216.
- 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.
- Brin, S., Motwani, R., Ullman, J. D., Tsur, S., 1997. Dynamic itemset counting and implication rules for market basket data. In: Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data. SIGMOD '97. ACM, New York, NY, USA, pp. 255–264.
- Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13 (1), 21–27.
- Fisher, R. A., 1938. The statistical utilization of multiple measurements. Annals of Eugenics 8 (4), 376–386.
- Hahsler, M., Grün, B., Hornik, K., 2005. arules - a computational environment for mining association rules and frequent item sets. Journal of Statistical Software, Articles 14 (15), 1–25.
- James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning with Applications in R. Springer.
- Johnson, I., 2017. arulesCBA: Classification Based on Association Rules. R package version 1.1.2.
- Johnson, R. A., Wichern, D. W., 2007. Applied Multivariate Statistical Analysis, Sixth Edition. Prentice Hall.
- Kuhn, M., 2017. caret: Classification and Regression Training. R package version 6.0-76.
- Lantz, B., 2013. Machine Learning with R. Packt.
- Liu, B., Hsu, W., Ma, Y., 1998. Integrating classification and association rule mining. In: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining. KDD'98. AAAI Press, pp. 80–86.
- Pérez, A., Larrañaga, P., Inza, I., 2009. Bayesian classifiers based on kernel density estimation: Flexible classifiers. International Journal of Approximate Reasoning 50 (2), 341–362, special Section on The Imprecise Dirichlet Model and Special Section on Bayesian Robustness (Issues in Imprecise Probability).

- SAS Institute Inc., 2011. The DISCRIM Procedure, SAS[®]/STAT User's Guide, Version 9.3. Cary, NC: SAS Institute Inc., 1974–2068.
- Venables, W. N., Ripley, B. D., 2002. Modern Applied Statistics with S, 4th Edition. Springer, New York, iISBN 0-387-95457-0.
- Weihs, C., Ligges, U., Luebke, K., Raabe, N., 2005. klaR Analyzing German Business Cycles. In: Data Analysis and Decision Support. Springer-Verlag, Berlin, pp. 335–343.