UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
Autorizada pelo Decreto Federal nº 77.496 de 27/04/76
Recredenciamento pelo Decreto nº17.228 de 25/11/2016
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA

XXVII SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS
SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA - 2023

COMPARAÇÃO DE ABORDAGENS DE APRENDIZADO DE MÁQUINA UTILIZANDO VARIÁVEIS CLIMÁTICAS DE ESTAÇÕES NA BAHIA

João Marcelo Souza Cézar Matos[1]; Maurício Santana Lordêlo[2]
1. Bolsista PROBIC, Graduando em Agronomia, Universidade Estadual de Feira de Santana, e-mail: jmarceloscm@gmail.com
2. Orientador, Departamento de Departamento de Ciências Exatas, Universidade Estadual de Feira de Santana, e-mail: mslordelo@uefs.br

PALAVRAS-CHAVE: aprendizado de máquina; climatologia; classificação

INTRODUÇÃO

Aprendizado de máquina é um campo de estudo que oferece aos computadores a capacidade de aprender sem serem explicitamente programados. Sendo assim, permite que o computador possa agir e tomar decisões a partir de dados (SAMUEL, 1959). Existem vários algoritmos, pertencentes ao grupo de aprendizagem supervisionada, que usam uma estrutura de árvore ramificada como base. Dentre eles, destacam-se a *Decision Tree* (Árvores de decisão) e *Random Forest* (Floresta Aleatória). A premissa básica de todos os algoritmos de classificação baseados em árvore é que por meio de uma sequência de perguntas, cada uma com resposta binária, as observações serão enviadas para uma ramificação (à esquerda ou à direita, dependendo de quais critérios eles atendem) separando-as em classes diferentes. Pode haver ramos dentro de ramos; e uma vez que o modelo é aprendido, ele pode ser representado graficamente como uma árvore (Rhys, 2020). A estrutura de árvore em formato de fluxograma tende a apresentar o resultado final em um formato legível, fornecendo uma visão de como e por que o modelo funciona ou não funciona bem para uma tarefa específica. Isso também torna as Árvores de Decisão particularmente apropriadas para aplicações nos quais o mecanismo de classificação precisa ser transparente por motivos legais, ou se os resultados precisam ser compartilhados com outros para informar as práticas de resultados futuros (Lantz, 2015).

O algoritmo KNN recebe esse nome pelo fato de usar informações sobre os k vizinhos mais próximos de uma observação para classificá-la em uma categoria. Definido esse número k, o algoritmo requer um conjunto de dados de treinamento composto de observações que foram classificadas e suas respectivas categorias. Então, para cada registro não rotulado no conjunto de dados de teste, KNN identifica os k registros nos dados de treinamento que são os" mais próximos" em similaridade.

Para aplicar os algoritmos a base de dados foi formada com a junção de 252 observações de cada estação, totalizando 756 observações. Foi adotada a mesma semente para o número aleatório e uma divisão em 70% para dados de treinamento e

30% para teste. Para as métricas de avaliações foram usadas a Acurácia e Coeficiente de *Kappa*.

MATERIAL E MÉTODOS OU METODOLOGIA (ou equivalente)

Os dados foram coletados na página do INMET (https://portal.inmet.gov.br). São 3 estações meteorológicas, localizadas em Salvador, Caravelas e Lençóis, com dados das variáveis climatológicas: Precipitação Total (mm), Pressão Atmosférica (mB), Temperatura Máxima (°C), Temperatura Média compensada (°C), Temperatura Mínima (°C), Umidade Relativa do Ar (%). A coleta dos dados foi realizada em estações meteorológicas convencionais com dados mensais e com registro de 2000 a 2020. Estações desse tipo exigem a presença diária de uma pessoa para coleta de dados.
Souza et al. (2012), com base na precipitação diária (mm), classificou a intensidade das chuvas nas seguintes categorias:

(1) dia "seco": abaixo de 2,2mm;
(2) chuva "muito fraca": entre 2,2 e 4,2mm;
(3) chuva "fraca": entre 4,2 e 8,4 mm;
(4) chuva "moderada": entre 8,4 e 18,6 mm;
(5) chuva "forte": entre 18,6 e 55,3 mm;
(6) chuva "muito forte": maior igual a 55,3 mm.

Três categorias foram utilizadas:
Abaixo de 4,2: Categoria A representa dia seco ou chuva muito fraca;
Entre 4,2 e 8,4: Categoria B representa chuva fraca;
Acima de 8,4: Categoria C representa chuva moderada, forte ou muito forte.

RESULTADOS E/OU DISCUSSÃO (ou Análise e discussão dos resultados)

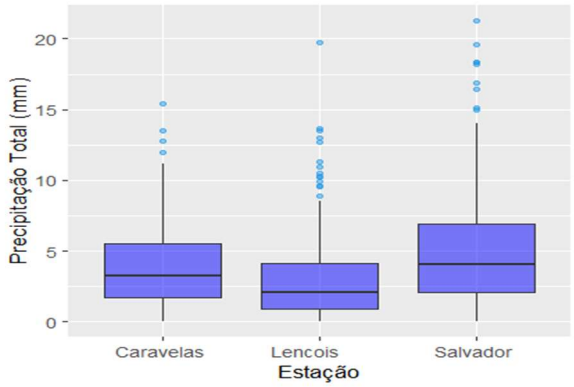
Os resultados das medidas descritivas para a base de dados completam com as três estações meteorológicas encontram-se na Tabela 1. As médias e as medianas foram iguais ou muito próximas para as medições das temperaturas. Como as variáveis possuem unidades de medidas diferentes, não é indicada a comparação da variabilidade das medições por meio do desvio padrão. Dessa forma, foi realizado o cálculo do coeficiente de variação, onde se observa que a temperatura mínima tem maior variabilidade e a pressão atmosférica a menor.

Tabela 1 – Medidas descritivas para junção dos dados das três estações meteorológicas

Medidas	Variáveis				
	Temperatura Max	Temperatura Med	Temperatura Min	Pressão ATM	Umidade Relativa
Mínimo	24,8	19,8	13,6	960,8	53,9
1 Quartil	27,7	23,7	19,4	967,0	75,9
Mediana	29,4	25,0	21,4	1009,0	80,2
Média	29,4	24,8	20,9	996,6	78,5
3 Quartil	30,8	26,1	22,7	1013,3	82,9
Máximo	35,3	28,6	25,8	1022,5	91,4
Desvio Padrão	1,9	1,7	2,3	22,7	6,8
Coef Variação (%)	6,7	6,8	11,1	2,3	8,7

A Figura 1, a seguir, apresenta medidas atípicas para as três estações. Além disso, mostra que a estação de Salvador apresentou maior dispersão dos dados e a de lençóis apresentou menor dispersão. Outra observação é com relação a maior amplitude dos dados na estação de Salvador**.**

Figura 1: Gráficos *Box-Plot* para Precipitação



A Tabela 2 e a Tabela 3, apresentadas a seguir, indicam os resultados das matrizes de confusão que apresenta o número de falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos para os algoritmos Árvore de Decisão e para o KNN.

Tabela 2 - Matriz de confusão para Árvores de decisão (*Decision Tree*)

Classificação predita	Classificação verdadeira		
	Categoria A	Categoria B	Categoria C
Categoria A	128	16	4
Categoria B	32	22	6
Categoria C	10	6	3

Tabela 3 – Matriz de confusão para K vizinhos mais próximos (KNN)

Classificação predita	Classificação verdadeira		
	Categoria A	Categoria B	Categoria C
Categoria A	132	13	3
Categoria B	34	21	5
Categoria C	10	4	5

Tabela 4 – Acurácia e *Kappa*

Algoritmo	Acurácia	*Kappa*
Decision Tree	0,67	0,28
KNN	0,69	0,31

Observando-se a Tabela 4, nota-se que o KNN apresentou maior número de acertos na matriz de confusão.

CONSIDERAÇÕES FINAIS (ou Conclusão)

A abordagem KNN obteve maior acurácia e também maior coeficiente Kappa quando comparado com a abordagem Árvore de decisão indicando ser o mais apropriado para a previsão dos resultados de precipitação pluviométrica classificada em três categorias de intensidade de chuva: A (dia seco ou chuva muito fraca), B (chuva fraca) e C (moderada, forte ou muito forte).

REFERÊNCIAS

LANTZ, B. 2015. Machine Learning with R: Expert techniques for predictive modeling to solve all your data analysis problems. Second Edition.

OLIVEIRA, H.L.C.O.; OLIVEIRA, S.R.M.; MONTEIRO, J.E.B.A. 2017. Geração de séries temporais de dados meteorológicos utilizando algoritmos de aprendizado de máquina. 11º Congresso Interinstitucional de Iniciação Científica. Campinas, São Paulo.

RHYS, H. I. 2020. Machine Learning with R, the tidyverse and mlr. Manning Publications Co.

SAMUEL, A. L. 1959. Some Studies in Machine Learning Using the Game of Checkers. Originally published in IBM Journal, Vol. 3, No.3.

SOUZA, W.M.; AZEVEDO, P.V.; ARAÚJO, L.E. 2012. Classificação da Precipitação Diária e Impactos Decorrentes dos Desastres Associados às Chuvas na Cidade do Recife-PE. Revista Brasileira de Geografia Física 02 250-268.