DADOS SINTÉTICOS APLICADOS EM ANÁLISES MULTIVARIADAS

Mylena Dias Delanora (PIBIC/CNPq/FA/UEM), Paulo Cesar Ossani (Orientador). E-mail: pcossani@uem.br.

Universidade Estadual de Maringá, Centro de Ciências Exatas, Maringá, PR.

Área e subárea do conhecimento: Probabilidade e Estatística, Probabilidade e Estatística Aplicadas.

Palavras-chave: Dados Sintéticos; Análise de Componentes Principais; Regressão Linear; Balanceamento de Classes.

RESUMO

Este trabalho avaliou o impacto das técnicas de geração de dados sintéticos SMOTE e ROSE na preservação da estrutura estatística de dados utilizados em análises multivariadas. Foram analisadas 13 bases de dados diferentes áreas do conhecimento, inicialmente desbalanceadas e posteriormente submetidas aos métodos de balanceamento. A avaliação foi realizada por meio da Análise de Componentes Principais (PCA) e de modelos de regressão linear ajustados às componentes principais. Os resultados mostraram que o SMOTE preservou de forma mais consistente a estrutura de variabilidade dos dados, mantendo autovalores e distribuição da variância próximos aos conjuntos originais. Em contraste, o ROSE promoveu alterações mais expressivas na distribuição da variância e apresentou desempenho mais heterogêneo na regressão linear. Em grande parte das bases, a primeira componente principal foi a principal variável explicativa da resposta, e o SMOTE proporcionou melhores coeficientes de determinação. Conclui-se que, embora ambas as técnicas sejam úteis para o balanceamento de dados, o SMOTE demonstrou maior capacidade de preservar as propriedades estatísticas dos conjuntos analisados, evidenciando a importância de considerar não apenas o desempenho preditivo, mas também a preservação estrutural dos dados.

INTRODUÇÃO

O crescente volume de dados gerados em diferentes áreas do conhecimento tem impulsionado o desenvolvimento de métodos estatísticos e computacionais voltados à extração de informações e ao apoio à tomada de decisões. Entretanto, a utilização de dados reais ainda enfrenta limitações relacionadas à privacidade, disponibilidade, custos de obtenção, desbalanceamento entre classes e presença de vieses, fatores que podem comprometer a qualidade das análises e o desenvolvimento de modelos de aprendizado de máquina (Boedihardjo, et al, 2021). Nesse contexto, os dados sintéticos surgem como uma alternativa para ampliar a disponibilidade de informações, preservar a confidencialidade dos dados originais e possibilitar estudos

em condições controladas (Reiter, 2005). Apesar dos avanços na geração de dados sintéticos, ainda permanecem desafios relacionados à preservação das características estatísticas dos dados originais, aspecto fundamental para garantir sua utilização em aplicações estatísticas. Dessa forma, este trabalho teve como objetivo avaliar o impacto da geração de dados sintéticos sobre a estrutura estatística de diferentes conjuntos de dados, por meio de técnicas de análise multivariada.

MATERIAIS E MÉTODOS

Foram utilizadas 13 bases de dados disponíveis no ambiente R, abrangendo diferentes áreas do conhecimento. As bases foram transformadas em problemas de classificação binária e submetidas a um processo de desbalanceamento artificial. Em seguida, aplicaram-se os métodos SMOTE (Chawla et al., 2002), que gera novas observações por interpolação entre amostras da classe minoritária, e ROSE (Menardi; Torelli, 2014), baseado na técnica de *smoothed bootstrap*. A avaliação foi realizada por meio da Análise de Componentes Principais (PCA) e de modelos de regressão linear, comparando os conjuntos de dados originais e balanceados quanto à preservação de suas características estatísticas. Também foi realizada a verificação da normalidade dos resíduos dos modelos de regressão por meio do teste de Shapiro–Wilk.

RESULTADOS E DISCUSSÃO

Os resultados da Análise de Componentes Principais (PCA) mostraram que o método SMOTE preservou a estrutura de variabilidade em 11 das 13 bases de dados analisadas, mantendo autovalores e proporções de variância explicada próximos aos observados nos dados originais. Em contrapartida, o método ROSE preservou adequadamente essa estrutura em 5 das 13 bases, enquanto nas demais promoveu redistribuições mais acentuadas da variância, geralmente concentrando-a nas primeiras componentes principais. Esses resultados indicam que o SMOTE apresentou maior capacidade de manter as características estatísticas dos conjuntos de dados, embora o desempenho de ambos os métodos tenha variado conforme as características de cada base analisada.

Para avaliar o desempenho dos modelos, foram ajustadas regressões lineares utilizando as componentes principais como variáveis explicativas. A Tabela 1 apresenta os coeficientes de determinação R^2, medida que indica a proporção da variabilidade da variável resposta explicada pelo modelo, para os dados originais e os conjuntos balanceados pelos métodos SMOTE e ROSE.

Tabela 1: Coeficientes de determinação dos dados.

BANCO DE DADOS	R^2 ORIGINAL	R^2 SMOTE	R^2 ROSE
USArrests	0,817	0,823	0,799
mtcars	0,74	0,787	0,802
swiss	0,546	0,613	0,594
airquality	0,46	0,558	0,384
longley	0,827	0,848	0,53
LifeCycleSaving	0,656	0,709	0,629
attitude	0,627	0,694	0,458
trees	0,775	0,792	0,744
ChickWeight	0,605	0,648	0,582
ToothGrowth	0,739	0,81	0,836
PlantGrowth	0,646	0,71	0,822
CO2	0,741	0,779	0,671
Orange	0,714	0,738	0,626

De modo geral, o SMOTE aumentou o coeficiente de determinação (R^2) nas 13 bases analisadas, enquanto o ROSE apresentou resultados mais variáveis. Além disso, a primeira componente principal foi a mais frequentemente significativa nos modelos, e o teste de *Shapiro–Wilk* indicou normalidade dos resíduos na maioria das bases, com algumas exceções.

CONCLUSÕES

Os resultados obtidos evidenciaram que as técnicas de geração e balanceamento de dados sintéticos exercem impactos distintos sobre a estrutura estatística dos conjuntos de dados. De modo geral, o método SMOTE preservou de forma mais consistente a estrutura multivariada dos dados e proporcionou melhores resultados nos modelos de regressão linear, enquanto o ROSE apresentou comportamento mais heterogêneo, promovendo alterações mais expressivas na distribuição da variância em algumas bases. Dessa forma, os resultados reforçam que a escolha da técnica de geração sintética deve considerar não apenas o desempenho preditivo, mas também a preservação das propriedades estatísticas dos dados, contribuindo para análises mais robustas e confiáveis.

AGRADECIMENTOS

Agradeço ao Programa PIBIC/CNPq/FA/UEM pelo projeto financiado e incentivo à pesquisa científica, que foram fundamentais para a realização deste trabalho. Estendo também meus agradecimentos ao professor Paulo César Ossani pela orientação, dedicação e pela oportunidade de aprendizado, que contribuíram significativamente para o desenvolvimento deste estudo.

REFERÊNCIAS

BOEDIHARDJO, M.; STROHMER, T.; VERSHYNIN, R. Covariance's loss is privacy's gain: computationally efficient, private and accurate synthetic data. **Foundations of Computational Mathematics**, 2022. Disponível em: https://doi.org/10.1007/s10208-022-09591-7. Acesso em: 2 set. 2025.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321-357, 2002. Disponível em: https://doi.org/10.1613/jair.953. Acesso em: 2 set. 2025.

MENARDI, G.; TORELLI, N. Training and assessing classification rules with imbalanced data. **Data Mining and Knowledge Discovery**, v. 28, n. 1, p. 92-122, 2014. Disponível em: https://doi.org/10.1007/s10618-012-0295-5. Acesso em: 2 set. 2025.

REITER, J. P. Using CART to generate partially synthetic public use microdata. **Journal of Official Statistics**, v. 21, n. 3, p. 441-462, 2005.