

Luiz Victor F. Santos

"Identificação das características mais relevantes para a priorização de correções de vulnerabilidades em softwares utilizando aprendizado de máquina"

=====

Pontos fortes:

- + Fácil de seguir. O artigo está muito bem escrito e fácil de ler.
- + Os autores utilizaram uma base de dados abrangente, contendo vulnerabilidades reportadas entre os anos de 1999 a 2021.
- + Seis algoritmos de classificação foram considerados no experimento.
- + Quatro algoritmos de Feature Selection foram usados pelos autores.
- + A utilização da plataforma Azure Machine Learning representa uma inovação significativa em relação a outras ferramentas mais antigas como o WEKA. O link abaixo mostra que a Azure Machine Learning (9th) está mais bem posicionada em Data Science em relação ao Weka (39th):

<https://6sense.com/tech/data-science-and-machine-learning/azuremachinelearning-vs-weka>

Pontos fracos:

- Pg. 7: Não foi mencionado no texto o que é a OffSec. Além disso, faltou uma nota de rodapé na página com o site oficial da OffSec.
- Pg. 7: O conceito de lematização (Lemmatization) foi explicado

de forma muito informal no texto. Eu sugiro adicionar algum artigo científico (e.g., [1]) que explique este conceito de maneira mais formal.

- Pg. 12: Achei a frase confusa: "as características que possuíam mais de 95%, 90%, 80%, 70%, 60% e 50% de valores zerados foram removidas".

Não seria algo assim? "as características que possuíam mais de 50% de valores zerados foram removidas".

- Em todo o texto há uma falta de padronização dos verbos na 3ª pessoa do singular. Exemplo: "permitindo", "utilizando" ao invés de "permitindo-se", "utilizando-se".

- Pg. 14: Como vocês garantem que o classificador Gradient Boosting não possui overfitting e que a acurácia dele é realmente alta (96,85%)? Foi usado validação cruzada para treinamento/teste do classificador? Se sim, quantos k-folds foram considerados?

- Pg. 18: Eu achei a afirmação "Essa diferença sugere que os métodos de correlação podem ser mais consistentes [...]" um pouco tendenciosa para ser colocada no artigo. Sugiro retirá-la. Eu acredito que seria necessário fazer um experimento dedicado envolvendo testes estatísticos/matemáticos para poder afirmar isso com propriedade. Ou pelo menos citar algum artigo científico que chegou nessa conclusão.

Referências Bibliográficas:

- [1] R. Pramana, Debora, J. J. Subroto, A. A. S. Gunawan and Anderies, "Systematic Literature Review of Stemming and Lemmatization Performance for Sentence Similarity," 2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA), Yogyakarta, Indonesia, 2022, pp. 1-6, doi: 10.1109/ICITDA55840.2022.9.