Ronaldo Mendonça Zica

"Estudo Empírico sobre Falhas em Projetos de Código Aberto de Aprendizado de Máquina"

==========

Pontos fortes:

+ Fácil de seguir: o texto está muito bem escrito e com poucos erros ortográficos.

+ Forte embasamento científico. O TCC cita o trabalho de Thomas Zimmermann [1] (pesquisador Sênior da Microsoft Research), que possui vários trabalhos relevantes na área de IA aplicada na Engenharia de Software, Produtividade no Desenvolvimento de Software, etc.

+ Amostra abrangente: O autor do trabalho analisou uma grande quantidade de projetos de Aprendizado de Máquina.

+ Boa visualização dos dados: O trabalho possui vários gráficos e tipos de gráficos que mostram com clareza os resultados obtidos.

+ Base de dados pública que pode ser utilizada por outros pesquisadores.

Pontos fracos:

- Alguns pontos do texto ficaram repetitivos. Exemplo na Pg. 27: "pesquisadores que se dedicaram ao estudo da evolução de códigos de Aprendizado de Máquina que possuem código aberto no Github". Sugiro alterar para: "pesquisadores que se dedicaram ao estudo da evolução de projetos de Aprendizado de Máquina que possuem código aberto no Github".

- Ainda na Pg. 27, foi mencionado sobre os projetos TensorFlow e Tesseract.

Todavia, o autor não citou nem mesmo o site dos dois projetos como nota de rodapé.

- Em vários pontos do texto (e.g., Pg. 30 da Tabela 1), ficou faltando os links para apontar para as referências bibliográficas.
- Pg. 33: Em todo o documento é usado ponto e vírgula ao invés da vírgula.
- Pg. 49-51: As Figuras 11, 12, e 13 usaram a média ao invés da mediana para o cálculo do fechamento de *issues* por projeto. Média e mediana não são similares.

Cuidado ao se utilizar a média em estudos científicos. A média aritmética não é um preditor confiável em Estatística. Ela apenas descreve o conjunto de dados como um todo e não diz o que está acontecendo dentro do conjunto. Por exemplo, a média é afetada por valores discrepantes (i.e., outliers).

A mediana é uma medida estatística fundamental na análise de dados. Ela é menos sensível a valores discrepantes do que a média, o que a torna mais robusta em certos contextos.

- Pg. 57: A expressão "outliers mais extremos" é um pleonasmo, assim como as expressões "subir para cima" ou "descer para baixo". Por definição, os outliers já são dados que se diferenciam drasticamente de todos os outros. São valores extremos que fogem da normalidade dos dados. Favor deixar somente a palavra "outliers".

- Pg. 59: Eu sugiro colocar também nas legendas das Figuras 21 e 24 que são os top-10 projetos mais populares do Github. Eu fiquei me perguntando aonde estavam os outros 92 projetos considerados no estudo.

Referências Bibliográficas:

[1] GONZALEZ, D.; ZIMMERMANN, T.; NAGAPPAN, N. The state of the ml-universe: 10 years of artificial intelligence & machine learning software development on github. p. 1–12, 2020. 22, 27, 33.