Nome do Aluno



**NOME DO TRABALHO**

CASCAVEL

MÊS / ANO

NOME DO ALUNO

**NOME DO TRABALHO**

Este trabalho é parte integrante da avaliação bimestral da disciplina de ……, do …º bimestre do curso de Engenharia de Software, xº Período.

Professor: Leandro de Souza

Cascavel, Ano

**Redimensionamento de dados**

O termo Big Data comumente usado de forma genérica, significa a enorme quantidade de dados que esta além da capacidade de processamento dos sistemas de gerenciamento de dados tradicionais devido a grande demanda de processamento para gerenciar e analisar os dados que possuem muitas formas e tipos.

Devido a essa característica frequente dos dados terem cada vez mais diferentes tipos de variáveis em um mesmo conjunto de dados e com um intervalo das variáveis muito diferentes entre si, usar uma escala original pode trazer deformação nos dados devido aos dados com maiores pesos nas variações.

Para solucionar esse problema é importante aplicar a redimensionamento na etapa de pré-processamento dos dados. Esse método que é dividida em duas técnicas, padronização e normalização, tem o objetivo de fazer com que dados com diferentes dimensões possam ser processados por um determinado algoritmo.

Em sua grande maioria, os algoritmos não conseguem trabalhar de forma eficiente com dados com diferentes pesos e dimensões e por isso é necessário padronizar ou normalizar, para isso o redimensionamento muda a escala dos dados para que esses dados para que possam estar em uma escala e possam ser compreendidos pelos algoritmos.

**Padronização e Normalização**

Técnicas aplicadas de forma frequente na etapa de preparação dos dados, são aplicados com o objetivo de colocar os dados em um intervalo de valor comum, essas técnicas não são obrigatórias, mas são recomendados para evitar que o algoritmo fique enviesado para as variáveis com grandezas maiores e torne a análise não confiável.

De forma geral, padronizar as variáveis vai resultar em uma média igual a 0 e um desvio padrão a 1, enquanto que na normalização as variáveis são inseridas em um intervalo de 0 e 1, podendo ser -1 e 1 caso exista resultado negativo.

Na técnica de normalização também conhecida como escala Mín-Max, faz com que os valores sejam deslocados e redimensionados para que fiquem entre 0 e 1 ou -1 e 1, aqui os outliers, que são valores extremos muito maiores ou menores em um determinado conjunto de dados, não são retirados.

Já a padronização, também conhecido como a normalização do escore Z, faz com que os valores sejam centralizados em torno da média com o desvio padrão da unidade, o que significa que a média do atributo se torna zero e o resultado da distribuição tem um desvio padrão da unidade.

**Quando usar a padronização e a normalização?**

A padronização pode ser útil quando os dados seguem uma distribuição gaussiana, seja, seguem uma distribuição normal com uma cursa simétrica que pode ser denominada Normal padrão ou Normal reduzidos. Do contrário da normalização, a padronização não tem um intervalo delimitador, dessa forma, mesmo que você tenha valores discrepantes em seus dados, eles não serão afetados pela padronização.

Já a normalização é uma opção melhor para momentos em que a distribuição dos dados não segue uma distribuição gaussiana, podendo ser útil em algoritmos que não assumem nenhuma distribuição de dados, assim como KNN e redes neurais.

# **Conclusão**

O grande volume de dados criado diariamente em todo o mundo trouxe grandes oportunidades para a área de análises que possui hoje uma gama maior de dados de diferentes formatos para a tomada de decisões, principalmente usados no universo empresário e de pesquisa.

Porem, esse grande volume de dados, o Big Data, criou a necessidade de existir métodos e ferramentas para o processamento, limpeza e demais etapas que devem ser seguidas para a utilização dos dados na real tomada de decisões.

Dentro do pré-processamento, temos o redimensionamento que é divido em duas técnicas que possuem o mesmo objetivo, transformar os dados para uma mesma escala e assim serem processados pelos algoritmos disponíveis.

Esse processo de redimensionar os dados, mesmo que não obrigatório, traz uma maior confiabilidade dos dados, já que as diferentes escalas dos dados, assim como os outliers podem trazer distorções nos resultados da distribuição em questão.

# 

# **Referências**

O que é Big Data. ETL Power Center. Disponível em: < [https://diogonvidal.wixsite.com/powercenter/post/o-que-%C3%A9-big-data](https://diogonvidal.wixsite.com/powercenter/post/o-que-é-big-data) >. Acesso em: 10 de setembro de 2022.

JUNIOR, Jose RF. Padronização vc. Normalização. Linkedin. Disponível em: < [https://www.linkedin.com/pulse/padroniza%C3%A7%C3%A3o-vs-normaliza%C3%A7%C3%A3o-jose-r-f-junior/?trk=pulse-article\_more-articles\_related-content-card&originalSubdomain=pt](https://www.linkedin.com/pulse/padronização-vs-normalização-jose-r-f-junior/?trk=pulse-article_more-articles_related-content-card&originalSubdomain=pt) >. Acesso em: 10 de setembro de 2022.

Distribuição Normal. IBICEL. Disponível em: < <https://www.ibilce.unesp.br/Home/Departamentos/CiencCompEstatistica/Adriana/a2-modelo-normal.pdf> >. Acesso em: 10 de setembro de 2022.

VAS, Arthur Lamblet. Normalização ou padronização as variáveis?. Data Hackers. Disponível em: < [https://medium.com/data-hackers/normalizar-ou-padronizar-as-vari%C3%A1veis-3b619876ccc9](https://medium.com/data-hackers/normalizar-ou-padronizar-as-variáveis-3b619876ccc9) >. Acesso em: 10 de setembro de 2022.

A importância da normalização e padronização dos dados em Machine Learning. IPNET Growth Partner. Disponível em: < <https://medium.com/ipnet-growth-partner/padronizacao-normalizacao-dados-machine-learning-f8f29246c12> >. Acesso em: 10 de setembro de 2022.

EDUARDO | Ciência dos Dados. ENTENDENDO DE VEZ A DIFERENÇA ENTRE NORMALIZAÇÃO E PADRONIZAÇÃO DOS DADOS. Youtube, data de publicação: 23 de out. de 2020. Disponível em: < <https://www.youtube.com/watch?v=-L8MacYRWGM> >. Acesso em: 10 de setembro de 2022