CENTRO NACIONAL DE MONITORAMENTO E ALERTAS
DE DESASTRES NATURAIS - CEMADEN

PROJETO DE PESQUISA FAPESP

APLICAÇÃO DE APRENDIZADO DE MÁQUINA PARA PREVISÃO DE ALAGAMENTOS

LINHA DE FOMENTO: BOLSA NO PAÍS - REGULAR - INICIAÇÃO CIENTÍFICA

Candidato:

Vitor Yuichi Hossaki

Orientador:

Prof. Dr. Leonardo Bacelar Lima Santos

São José dos Campos, SP

30 de setembro de 2021

Sumário

1 Introdução

Os alagamentos são fenômenos cada vez mais frequentes em regiões urbanas devido ao aumento da população e o crescimento desordenado do processo de urbanização. Em São Paulo, os alagamentos são recorrentes desde os primórdios de sua ocupação, a estrutura urbana aliado às características dos rios existentes auxiliam na deflagração destes fenômenos [Hirata et al., 2013]. Segundo [Santos, 2013], estima-se que os efeitos macroeconômicos dos alagamentos são de 172.3 milhões de reais por ano, afetando setores logísticos e industriais.

Diante dos impactos materiais, econômicos e humanos causados pelos alagamentos, é necessário medidas que visem a mitigação e antecipação deste fenômeno. Estas medidas estão associadas a uma infinidade de maneiras como os sistemas de alertas, essencial para que a comunidade seja alertada com antecedência de fenômenos naturais intensos e, desta forma, minimizar e previnir possíveis danos materiais e humanos [Kobiyama et al., 2006].

Nesse ínterim, algumas pesquisas como de [Horita et al., 2015] e [Hirata et al., 2013], demonstram que a utilização de redes sociais que contém informações geográficas voluntárias podem ser um instrumento efetivo para o desenvolvimento de sistemas monitoramento e alertas das possíveis ocorrências de alagamento. A finalidade deste projeto é definir qual é o algoritmo de aprendizado de máquina que possui maior precisão com relação à previsão de alagamentos sobre determinado conjunto atributos.

2 Objetivos

A finalidade deste projeto é alinhar-se com os objetivos do Centro Nacional de Alerta de Desastres Naturais (CEMADEN) no desenvolvimento de meios para prever alagamentos, partindo-se da hipótese que é possível utilizar algoritmos de aprendizado de máquina para detectar alagamentos, e desta forma, emitindo-se alertas.

Objetivos específicos:

1. Definir qual é o algoritmo mais preciso para relacionar os dados metereológicos, pluviométricos e frequência de tweets para o modelo de classificação de dias de alagamentos e não-alagamentos.

3 Justificativa e relevância do projeto

O desenvolvimento acelerado de São Paulo culminou na urbanização descontrolada causando diversas consequências na região. A impermeabilização do solo, a drenagem urbana deficitária e a topografia favorável ao acúmulo de água, são reflexos desta expansão desordenada. A deflagração de alagamentos causaram diversas perdas diretas e indiretas para o PIB de São Paulo, alcançando média de 172 milhões de reais em prejuízos econômicos por ano em algumas regiões do estado. Diante dos danos humanos, materiais e econômicos que os alagamentos vêm causando ao longo das décadas, é necessário medidas mitigadoras e o desenvolvimento de sistemas de alertas que possam antecipar a possível deflagração do fenômeno hidrológico.

Além disso, em virtude da ascensão meteórica da tecnologia e da influência das redes sociais, o projeto visa o desenvolvimento da computação aplicada aos fenômenos hidrológicos de alagamentos, utilizando-se conceitos de ciência de dados e *Machine Learning*.

4 Fundamentação Teórica

4.1 Alagamentos

Os alagamentos são fenômenos associados ao acúmulo de água em determinado local, favorecidos pela microdrenagem e macrodrenagem insuficientes. A ausência de planejamento urbano e a rápida modificação do espaço culmina na impermeabilização do solo, contribuindo para diminuição do tempo concentração e o aumento do volume de escoamento superficial, amplificando-se assim, o pico da vazão e consequentemente saturando a drenagem pluvial do local [Hansmann, 2013].

A topografia e a elevação do local também são fatores preponderantes para a ocorrência de alagamentos, ou seja, verifica-se que os lugares com maior frequência de alagamentos tem características morfométricas planas, depressões ou fundos de vales, dificultanto o processo de escoamento superficial do local [Braga, 2016].

Fatores como o descarte inadequado de resíduos sólidos, podem causar obstrução dos sistema de drenagem do local, isto ocorre em decorrência da ausência de educação ambiental da população.

Os alagamentos são fenômenos complexos, uma vez que a sua causa está interrelacionada a uma gama parâmetros como o clima que incluem precipitação forte ou persistente, falhas estruturais urbanas, sistemas de drenagem inadequados, bacias hidrográficas, proximidades de corpos aquáticos, uso

e ocupação inapropriado do solo e entre outros [Doocy et al., 2013].

4.2 A utilização de redes sociais para o monitoramento de eventos

O desenvolvimento da sociedade na esfera tecnológica permitiu a ascensão meteórica das redes sociais e suas funcionalidades. A quantidade massiva de dados gerados das redes sociais consolidam a interação do universo virtual com o mundo concreto, onde usuários expressam suas percepções e emoções acerca dos eventos circundantes [Naaman, 2011]. A atividade das redes sociais e sua heterogeneidade espacial demonstra a potencialidade para o monitoramento de eventos metereológicos como a precipitação [de Andrade et al., 2021].

Através das plataformas de mídia social, uma única postagem pode ser vista por milhares de usuários simultaneamente, além disso algumas plataformas utilizam-se de georreferenciamento que permite a visualização não só da postagem como também a localização do usuário com seu dispositivo móvel. Redes sociais (Twitter) ou aplicativos como Open Street Map que permitem a tecnologia de georreferenciamento são denominadas informações geográficas voluntárias, o trabalho de [Horita et al., 2015], integra estas plataformas para o gerenciamento de risco dos alagamentos.

A utilização das redes sociais apresentam uma crescente tendência na sua incorporação em pesquisas para o monitoramento e análise de uma infinidade de eventos. Segundo [De Albuquerque et al., 2015], a utilização de informações geográficas voluntárias, principalmente a rede a social Twitter, são componentes fundamentais para a maior consciência dos eventos ocorrentes ou seja, consolida-se a percepção dos elementos no ambiente e possibilita maior compreensão das possíveis consequências.

4.3 Classificação

O Aprendizado de Máquina é cada vez mais empregado pelos pesquisadores na área de desastres de naturais, alguns autores utilizam esta ferramenta para analisar a semântica atrelada das postagens de rede social, e assim, aprimorar os resultados da classificação de determinada ocorrência [De Albuquerque et al., 2015, Deparday et al., 2019].

Esta tecnologia pode ser definida como um conjunto métodos computacionais para aprimorar performance ou realizar predições acuradas. A classificação é um dos métodos computacionais amplamente utilizados para categorização de cada item em uma série de dados. Matematicamente, a classificação é descrita por uma função $F : \mathcal{X} \longrightarrow \mathcal{Y}$ que associa elementos no conjunto de atribu-

tos $\mathcal{X}$ a uma classe de $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$, com $n \in \mathbb{N}^*$, e partindo-se de um indicador de classe $\mathcal{Y} = \{1, 2, ..., n\}$, portanto, quando $x \in \mathcal{X}$ e $y \in \mathcal{Y}$, a função $y = F(x)$ indica que x pertence à Ω_y.

Os modelos de aprendizagem supervisionada, a função F utiliza-se das informações do conjunto de treinamento representado pela equação $\mathcal{D} = \{(x_j, \omega_j \in \times\Omega : i = 1, ...m;\ j = 1, ..., c\}$, no qual m é a quantidade de dados no treinamento.

Atualmente os principais algoritmos empregados para classificação são o *Suport Vector Machine* (SVMs) e *Random Forest* [Mohri et al., 2018].

4.3.1 Suport Vector Machines (SVMs)

O método SVM realiza a distinção entre amostras de treinamento partindo-se de um hiperplano que possui maior abrangência de separação, mapeando o padrão de vetores para um espaço de alta dimensão, determinando-se o hiperplano mais adequado para separação de dados. Este algoritmo é utilizado por diversos autores devido à alta acurácia para problemas de classificação binária [Lian and Lu, 2006] .

O hiperplano corresponde ao lugar geométrico nos quais a função $f(x) = \langle w, x\rangle + b$ é nula. A variável w é o vetor ortogonal ao hiperplano e b a distância entre a função e a origem do espaço de atributos.

Para se encontrar o hiperplano mais adequado para separação entre as classes, é necessário a resolução do problema de otimização [Theodoridis et al., 2010] representado por: $max_\gamma(\sum_{i=1}^{m}\gamma_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\gamma_i\gamma_j y_i y_j\langle x_i, x_j\rangle)$, $\begin{cases} 0 \leq \gamma_i \leq \mathcal{C}, i = 1, ..., m \\ \sum_{i=1}^{m}\gamma_i y_i = 0 \end{cases}$, a variável $\mathcal{C}$ é o parâmetro utilizado para regularização para ajustar o hiperplano e γ_i são os multiplicadores de Lagrange.

A definição dos parâmetros w e b que compõem o hiperplano são dadas por: $w = \sum_{\forall x_i \in SV} y_i\gamma_i x_i$, $b = \frac{1}{\#SV}(\sum_{x_i \in SV} y_i + \sum_{x_i \in SV} \cdot \sum_{x_j \in SV}\gamma_i\gamma_j y_i y_j\langle x_i, x_j\rangle)$, SV é um subconjunto das amostrar de treinamento $\mathcal{D}$, nos quais os elementos são os vetores suporte. Por fim indicação da classe pertencente do vetor analisado é dado pelo sinal da função discriminante $f(x)$ [Maselli and Negri, 2019].

4.3.2 Floresta Aleatória (*Random Forest*)

A classificação através do algoritmo Floresta Aleatória vem sendo amplamente utilizada literatura para avaliação e mapeamento dos padrões de eventos hidrológicos. Pesquisa como de [Zhu and

Zhang, 2021] e [Liu et al., 2020] demonstram a potencialidade do algoritmo para avaliar a resiliência e os padrões espaciais dos alagamentos.

Este modelo é um algoritmo de classificação que representa um conjunto de árvores de decisão, que combina a saída destas diversas árvores atribuindo-se uma classe ao conjunto de dados. Segundo [Breiman, 2001], a Floresta Aleatória consiste em uma coleção de classificadores em forma de árvore descritos por $\{h(x, \theta_k), k = 1, ..\}$ onde θ_k são independentes e em cada árvore é lançado um voto unitário para a classe mais popular para o input x.

4.4 Redes Neurais

Este algoritmo vem sendo empregado para emissão de alertas hidrológicos e mapeamentos de suscetibilidade em alguns autores como [da Silva et al., 2016] e [Pacheco Quevedo et al., 2020], demonstrando efetividade e acurácia elevada para os modelo de previssão associados aos fenômenos hidrológicos.

A técnica *Multilayer Perceptron* demonstra resultados relevantes as mais diversas áreas da ciência [Gardner and Dorling, 1998]. Este algoritmo consiste em um sistema interconectado de neurônios, estes nós são conectados entre si por um peso. Matematicamente as camadas de neurônios de entrada e saída são vetores definidos como i e O respectivamente, e os pesos como uma matriz W. Portanto a saída da rede é dada por $O = f(IW_{io})$, ao final do processo uma função determina se aquele nó será ativado na condição $f(x) \begin{cases} 1 \;\; x > 0 \\ \\ 0 \;\; otherwise \end{cases}$. Assumindo que T é o parâmetro de saída para o vetor de treinamento, o algoritmo o calcula o erro associado através de $E(O) = T - O = T - f(IW_{io})$. Algumas técnicas visam a redução do erro através da atualização dos pesos no processo representado matematicamente por $W_{io}(t + 1) = W_{io}(t) + \alpha E_n$.

5 MATERIAIS E MÉTODOS

5.1 Área de estudo e dados disponíveis

O estudo foi realizado na região de São Paulo onde está localizado a bacia hidrográfica do Rio Tamanduateí (Figura 1). Esta bacia possui uma área de $323km^2$ e se estende até as bacia hidrográficas do Rio Pinheiro, Rio Guaió, Rio Aricanduva e Córrego de Tapuapé. Nesta região, foi analisado a

partir de um pluviômetro um raio espacial de $2000m$ que abrange as regiões de alagamentos, tweets georrefenciados e a célula de radar.

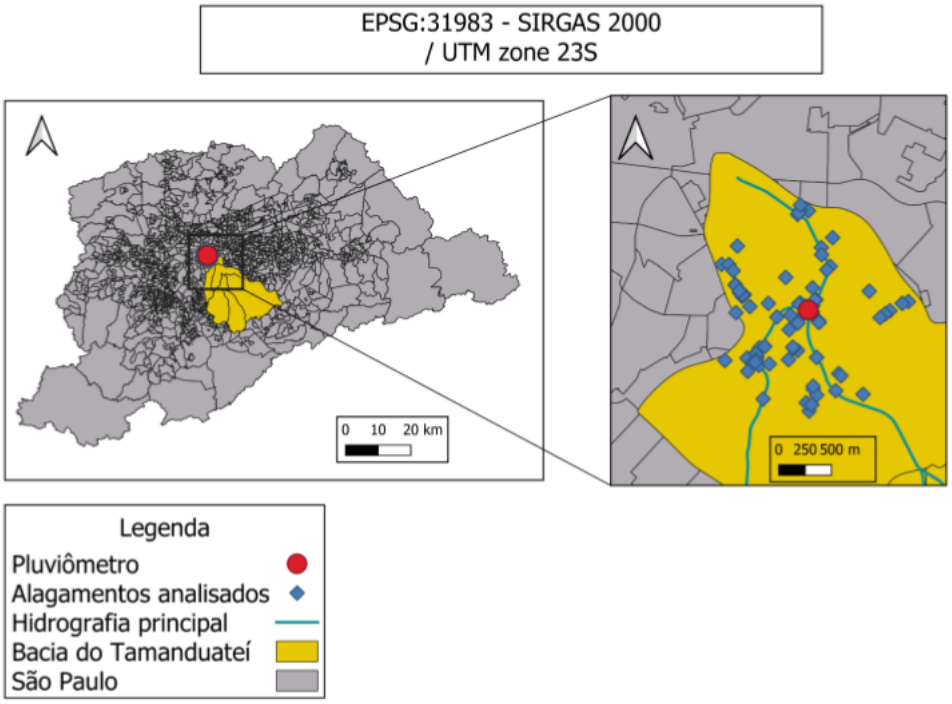

Figura 1: Área de estudo

Os dados da rede social Twitter foram extraídas através da API (*Application Programming Interface*). Os dados pluviométricos são coletados do pluviômetro 833A, pertencente ao Centro Nacional e Alertas e Desastres Naturais (CEMADEN), estes dados podem ser encontrados no próprio site da instituição.

A série histórica de alagamentos na área de estudo, foram concebidas por um dos integrantes da pesquisa. Os dados metereológicos foram extraídos por estações pertencentes ao CEMADEN, o equipamento está localizado na cidade de São Roque - SP e atualmente está em operação pelo Departamento de Controle do Espaço Aéreo (DECEA). Esse radar tem alcance de 250 km, cobrindo toda a região metropolitana de São Paulo. O produto de radar usado para o CAPPI (Constant Altitude Plan Position Indicator) na altura de 3 km. Este produto possui uma resolução espacial de aproximadamente 1 km e uma resolução temporal de 10 minutos. Para a conversão da refletividade (dBZ) em taxa de separação (mm / h) foi utilizado em relação a Marshall-Palmer [Marshall andW, 1948]) e a seguir os dados foram acumulados por dia.

5.2 Ferramentas

A análise e aplicação do projeto será realizada de maneira geral com a ferramenta *Python*. Para a manipulação, filtragem e tratamento dos dados será utilizada a biblioteca *Pandas*, já a análise gráfica com *Matplotlib* e *Seaborn*.

A aplicação de testes estatísticos na série de dados será usado *Scipy* e *Numpy*, para os modelos de apredizagem supracitados, a biblioteca específica para aprendizado de máquina denominado *Scikit-learn*. Por fim, algumas filtrações no banco de dados de alagamentos será realizada com ferramentas de geoprocessamento do software *QGIS*.

5.3 Proposta de algoritmo para definição do modelo

A concepção inicial deste trabalho é analisar as séries temporais dos alagamentos, tweets, pluviômetro e radar, para definir quais são o melhor conjuntos de parâmetros em dias de alagamentos, associando-se ao número mínimo necessário de tweets para emissão de um alerta.

A série temporal analisada compreende os três primeiros meses do ano de 2019. Para a base de dados dos tweets, o processamento consiste no recorte temporal e filtração do tweets com base na lista de palavras associadas ao contexto metereológico e hidrológico. Esta lista de palavra basea-se no trabalho de [de Andrade et al., 2021].

Com base na estrutura (Figura 2), será registrado em único arquivo, na mesma série temporal, o número de tweets filtrados, os valores de precipitação do radar e o pluviomêtro, e se houve alagamentos no dias analisados. Este dados processados em um único arquivo, possibilitarão a submissão nos modelos de aprendizados propostos, dividindo-se em base de dados para teste e treinamento.

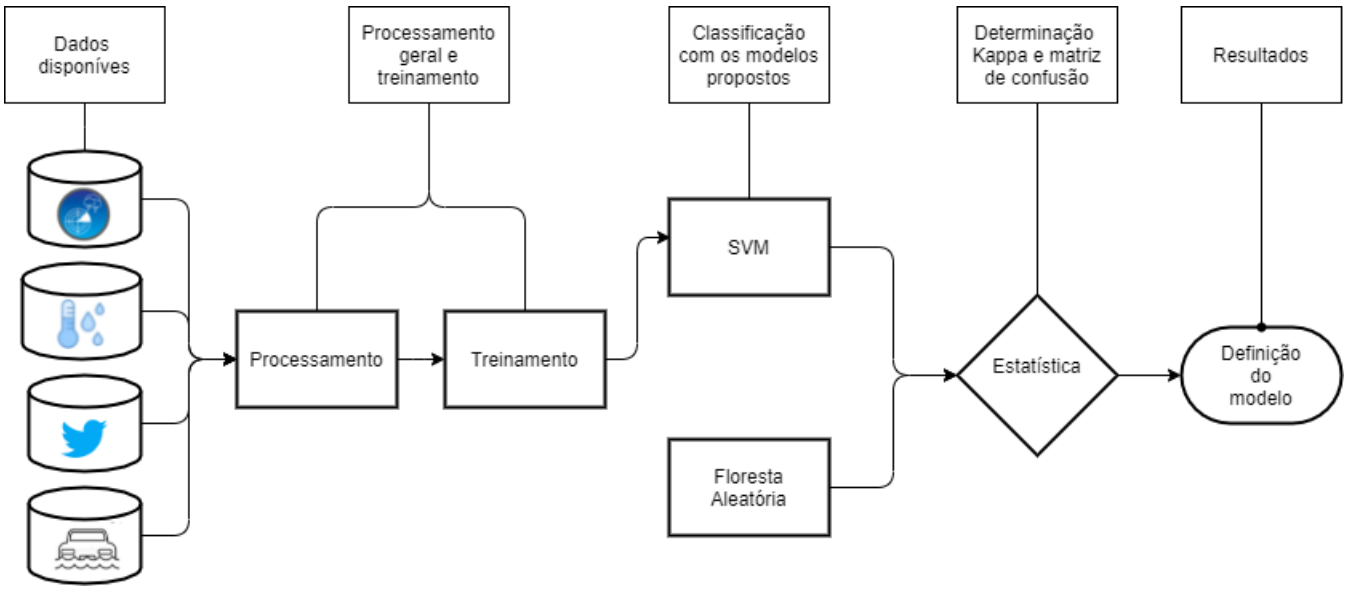

Figura 2: Metodologia

Como a classificação binária consiste em dias de alagamento e não alagamento, a acurácia será medida a partir da base de dados de teste. Após o treinamento nos modelo SVM e Floresta Aleatória e Redes Neurais, serão analisadas a acurácia através da validação cruzada e subsequentemente testes estatísticos como ANOVA e coeficiente Kappa, determinando-se assim, o algoritmo que possui maior potencial para o desenvolvimento de um sistema de alerta com base nos dados disponíveis.

6 Cronograma

A pesquisa será realizada em 12 meses e será executada nos passos listado abaixo (Tabela 1)

A - Revisão sistemática em desastres associados à alagamentos e modelos de classificação;

B - Estudo dos modelos de aprendizado de máquina e aplicação em Python;

C - Processamento dos bancos de dados;

D - Análise exploratória dos dados processados;

E - Submissão dos dados processados para treinamento nos modelos propostos;

F - Classificação;

G - Cálculos estatísticos e inferências;

H - Alterações, ajustes e otimizações no modelo de melhor desempenho;

I - Análise e conclusão dos resultados;

J - Relatório final

Mês		1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º	12º
Etapas	A	●	●										
	B	●	●	●									
	C			●	●								
	D				●								
	E					●	●						
	F					●	●						
	G							●	●				
	H								●	●			
	I										●	●	
	J											●	●

Tabela 1: Cronograma

Referências

[Braga, 2016] Braga, J. O. (2016). Alagamentos e inundações em áreas urbanas: estudo de caso na cidade de santa maria-df.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[da Silva et al., 2016] da Silva, M. R., Santos, L. B. L., Scofield, G. B., and Cortivo, F. D. (2016). Utilização de redes neurais artificiais em alertas hidrológicos: Estudo de caso na bacia do rio claro em caraguatatuba, estado de são paulo. *Anuário do Instituto de Geociências*, 39(1):23–31.

[De Albuquerque et al., 2015] De Albuquerque, J. P., Herfort, B., Brenning, A., and Zipf, A. (2015). A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International journal of geographical information science*, 29(4):667–689.

[de Andrade et al., 2021] de Andrade, S. C., Porto de Albuquerque, J., Restrepo-Estrada, C., Westerholt, R., Rodriguez, C. A. M., Mendiondo, E. M., and Delbem, A. C. B. (2021). The effect of intra-urban mobility flows on the spatial heterogeneity of social media activity: investigating the response to rainfall events. *International Journal of Geographical Information Science*, pages 1–26.

[Deparday et al., 2019] Deparday, V., Gevaert, C. M., Molinario, G., Soden, R., and Balog-Way, S. (2019). Machine learning for disaster risk management.

[Doocy et al., 2013] Doocy, S., Daniels, A., Murray, S., and Kirsch, T. D. (2013). The human impact of floods: a historical review of events 1980-2009 and systematic literature review. *PLoS currents*, 5.

[Gardner and Dorling, 1998] Gardner, M. W. and Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636.

[Hansmann, 2013] Hansmann, H. Z. (2013). Descrição e caracterização das principais enchentes e alagamentos de pelotas-rs. *Universidade Federal de Pelotas, Pelotas-RS*.

[Hirata et al., 2013] Hirata, E., Giannotti, M. A., Larocca, A. P. C., and Quintanilha, J. A. (2013). Mapeamento dinâmico e colaborativo de alagamentos na cidade de são paulo. *Boletim de Ciências Geodésicas*, 19:602–623.

[Horita et al., 2015] Horita, F. E., de Albuquerque, J. P., Degrossi, L. C., Mendiondo, E. M., and Ueyama, J. (2015). Development of a spatial decision support system for flood risk management in brazil that combines volunteered geographic information with wireless sensor networks. *Computers & Geosciences*, 80:84–94.

[Kobiyama et al., 2006] Kobiyama, M., Mendonça, M., Moreno, D. A., Marcelino, I., Marcelino, E. V., Gonçalves, E. F., Brazetti, L. L., Goerl, R. F., Molleri, G. S., and Rudorff, F. d. M. (2006). *Prevenção de desastres naturais: conceitos básicos*. Organic Trading Curitiba.

[Lian and Lu, 2006] Lian, H.-C. and Lu, B.-L. (2006). Multi-view gender classification using local binary patterns and support vector machines. In *International Symposium on Neural Networks*, pages 202–209. Springer.

[Liu et al., 2020] Liu, D., Fan, Z., Fu, Q., Li, M., Faiz, M. A., Ali, S., Li, T., Zhang, L., and Khan, M. I. (2020). Random forest regression evaluation model of regional flood disaster resilience based on the whale optimization algorithm. *Journal of Cleaner Production*, 250:119468.

[Marshall andW, 1948] Marshall andW, J. (1948). Mc. k. palmer,"the distribution of raindrops with size,". *J. Meteor*, 5:165–166.

[Maselli and Negri, 2019] Maselli, L. Z. and Negri, R. G. (2019). Integração entre estratégias multi-classes e diferentes funções kernel em máquinas de vetores suporte para classificação de imagens de sensoriamento remoto. *Revista Brasileira de Cartografia*, 71(1):149–175.

[Mohri et al., 2018] Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

[Naaman, 2011] Naaman, M. (2011). Geographic information from georeferenced social media data. *SIGSPATIAL Special*, 3(2):54–61.

[Pacheco Quevedo et al., 2020] Pacheco Quevedo, R., Garcia de Oliveira, G., and Antonio Guasselli, L. (2020). Mapeamento de suscetibilidade a movimentos de massa a partir de redes neurais artificiais. *Anuario do Instituto de Geociencias*, 43(2).

[Santos, 2013] Santos, E. T. d. (2013). *Impactos econômicos de desastres naturais em megacidades: o caso dos alagamentos em São Paulo*. PhD thesis, Universidade de São Paulo.

[Theodoridis et al., 2010] Theodoridis, S., Pikrakis, A., Koutroumbas, K., and Cavouras, D. (2010). *Introduction to pattern recognition: a matlab approach*. Academic Press.

[Zhu and Zhang, 2021] Zhu, Z. and Zhang, Y. (2021). Flood disaster risk assessment based on random forest algorithm. *Neural Computing and Applications*, pages 1–13.