Cemaden
Centro Nacional de Monitoramento
e Alertas de Desastres Naturais

FLOODING FORECAST VIA INTEGRATION OF TWEETS, WEATHER DATA AND MACHINE LEARNING ALGORITHMS

Student: Vitor Yuichi Hossaki
Supervisor: Prof. Dr. Leonardo Bacelar Lima Santos

Report to CNPq/Cemaden Scientific Initiation Scholarship Program

December of 2021

SUMÁRIO

Pág.

1 INTRODUCTION

Flooding is a frequent phenomenon in urban regions due to population growth and the characteristics of the urbanization process, causing an increase of impervious surfaces and poor drainage. This hydrological phenomenon is among the natural hazard associated with the most significant impact in the world. Over the years,

there has been an increase in its global incidence. Global factors, such as Global warming, and local ones, such as the lack of urban planning (TINGSANCHALI, 2012) are some of the variables causing this increase.

In the city of São Paulo, Brazil, flooding has been recurrent since the beginning of its occupation. The urban structure combined with the hydrographic and morphological characteristics helps trigger this phenomenon (HIRATA et al., 2013). Santos (2013) estimated that the macroeconomic effects of flooding are 172.3 million reais per year. Based on the impacts mentioned, there is a growing trend in the literature incorporating tools such as the social network to predict flooding.

In this context, some researchers, such as those presented by Horita et al. (2015) and Hirata et al. (2013), demonstrate that the use of social networks provided with voluntary geographic information can be used as an effective instrument in the development of flood monitoring and warning systems.

Similarly, Albuquerque et al. (2015) demonstrates the potential of spatiotemporal data obtained through the social network Twitter for disaster management. In the cited search, posts are filtered through keywords and then binary sorted among those "related" or "not related" to the event of interest, which in turn are aligned to a time series of outbreaks of flooding in a given region. The results obtained indicate the statistical relevance of the information pointed out by the Tweets in relation to the flooded regions. In a similar way Hossaki et al. (2021b) also demonstrates that the words used in the Twitter social network linked to the meteorological context on days of flooding are greater than in relation to days that do not occur.

In time, it is worth highlighting the recent rise in the use of Artificial Intelligence techniques, mainly Machine Learning methods, in the management and decision-making in the face of disaster events. In Sit et al. (2019) natural language processing algorithms are used for semantic identification of Tweets related to the context of

disasters. Machine Learning is also incorporated into the analysis of multiple physical parameters useful in flood prediction, as shown by Mosavi et al. (2018).

Thus, in the context of the problem involving flooding events and the potential offered by Machine Learning methods, this research project aims to build an algorithm capable of predicting the occurrence of flooding through information automatically extracted from the social network. Twitter. Data obtained by meteorological radar, rain gauge and the flooding database are components that integrate the proposed algorithm.

2 OBJECTIVES

This research project has as main objective the development of a method for flood forecasting based on Machine Learning techniques, temporal data extracted from Tweets, meteorological radar and rain gauges.

In addition, more specific goals are:

a) Analyze and define the best performing Machine Learning algorithm for flood prediction based on data extracted from Tweets and meteorological instruments (weather radar and rain gauges);

b) Investigate the attributes (i.e., data from Tweets, weather radar, rain gauges and/or derivatives) of greatest relevance for flood prediction;

c) Structure and make available in a public repository the databases used and source codes implemented in the research;

d) Disseminate the results obtained at scientific events and journals.

3 THEORY FUNDAMENTALS

3.1 Floods

Flooding is a complex phenomenon, as its cause is interrelated to a range of parameters such as climate, urban structural faults, inadequate drainage systems, hydrographic basins, proximity to water bodies, inappropriate use and occupation of the soil, among others (DOOCY et al., 2013).

The absence of urban planning and the rapid modification of the space culminate in soil sealing, contributing to a reduction in concentration time and an increase in the volume of surface runoff, thus amplifying the peak flow and consequently saturating the site's rainwater drainage (HANSMANN, 2013).

The local topography is also a preponderant factor for the occurrence of flooding, since it has already been verified that the places with the highest frequency of flooding have flat morphometric characteristics, depressions or valley bottoms, hindering the local runoff process (BRAGA, 2016).

Also, due to the population's lack of environmental education, the inadequate disposal of solid waste appears as another factor causing the obstruction of the local drainage system, once again leading to flooding.

3.2 Social media for flood monitoring

The development of society in the technological sphere allowed the meteoric rise of social networks and their functionalities. The massive amount of data generated by these networks consolidates the interaction of the virtual universe with the concrete world, where users express their perceptions and emotions about the surrounding events (NAAMAN, 2011).

The activity of social networks and their spatial heterogeneity demonstrate the po-

tential for monitoring meteorological events (ANDRADE et al., 2021). Meanwhile, the work of Horita et al. (2015) integrates these platforms for flood risk management.

The use of social networks shows a growing trend regarding their incorporation in research for the monitoring and analysis of different types of events. According to Albuquerque et al. (2015), the use of voluntary geographic information, mainly from the Twitter network, is a fundamental component for greater awareness of disaster events. Thus consolidating a better perception of the environment, in addition to enabling greater understanding of the possible consequences of phenomena such as precipitation (HOSSAKI et al., 2021a).

3.3 Classification

Machine Learning Techniques are increasingly employed in the study of natural disasters. Some researches use such techniques to analyze the semantics linked to posts on social networks, thus allowing to improve the classification results of a certain type of event (ALBUQUERQUE et al., 2015; DEPARDAY et al., 2019).

Classification comprises the widely used methods for associating each item in a data series with a particular class. Formally, the classification is described by a function $F : \mathcal{X} \rightarrow \mathcal{Y}$ that associates elements in the attribute set $\mathcal{X}$ to a class of $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$, with $n \in \mathbb{N}^*$ through a class indicator $\mathcal{Y} \in \{1, 2, ..., n\}$. Under these conditions, when $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, the function $y = F(x)$ indicates that x belongs to Ω_y. Regarding the supervised learning classification models, the function F makes use of information extracted from the training set $\mathcal{D} = \{(x_j, \omega_j) \in \mathcal{X} \times \Omega : i = 1, ..., m\}$.

Among several existing proposals in the literature, *Support Vector Machine* (SVMs), *Random Forest* (RF) and *Multilayer Perceptron* (MLP) are frequently used in the most diverse application domains.

3.3.1 SVM

The SVM method distinguishes between training examples based on a hyperplane with greater margin of separation, either in the original data space or conveniently remapped. According to Lian e Lu (2006), this method is used by several authors due to its high accuracy and generalizability.

The hyperplane corresponds to the locus where $f(x) = \langle w, x\rangle + b$ is null. The variable w is the vector orthogonal to the hyperplane and b the distance between the hyperplane and the origin of the attribute space. The determination of the hyperplane with the largest separation margin is obtained by optimizing the (THEODORIDIS et al., 2010) problem:

$$\begin{aligned} &\max_{\gamma}(\textstyle\sum_{i=1}^{m}\gamma_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\gamma_i\gamma_j y_i y_j\ langle x_i, x_j\rangle) \\ &\text{subject to } \begin{cases} 0 \leq \gamma_i \leq C, i = 1, ..., m \\ \sum_{i=1}^{m}\gamma_i y_i = 0 \end{cases} \end{aligned} \tag{3.1}$$

where C is the parameter used to regularize the hyperplane and γ_i are Lagrange multipliers.

The definition of the parameters that determine the hyperplane are $w = \sum_{\forall x_i \in SV} y_i\gamma_i x_i$ and $b = \frac{1}{\#SV}(\sum_{x_i \in SV} y_i + \sum_{x_i \in SV} \cdot \sum_{x_j \in SV} \gamma_i\gamma_j y_i y_j \langle x_i, x_j\rangle)$, where SV is a subset of the samples in $\mathcal{D}$ such that $\gamma_i \neq 0$, denoted by support vectors. Finally, the indication of the class associated with the analyzed vector is given by the sign of the discriminant function $f(x)$ (MASELLI; NEGRI, 2019).

3.3.2 RF

The RF method has been widely used in applications related to the identification of hydrological events. In Zhu e Zhang (2021) and Liu et al. (2020) the potential of this method in evaluating the resilience and identifying spatial patterns of flooding

is demonstrated. In a superficial way, according to Breiman (2001), this method is represented by a set of decision trees that combine their respective outputs through a majority voting scheme in order to make a final decision.

3.3.3 MLP

The MLP method has been used to issue hydrological alerts and susceptibility mappings (SILVA et al., 2016; QUEVEDO et al., 2020).

This method comprises a system of weighted connections between artificial neurons distributed in different layers. Mathematically the layers of input and output neurons are vectors $\mathbf{i}$ and $\mathbf{o}$, respectively, and the weights structured in a matrix $\mathbf{W}$. Under these conditions, the output of the network is given by $\mathbf{o} = f(\mathbf{iW})$, where $f(x)\begin{cases} 1, & x > 0 \\ 0, & \text{otherwise} \end{cases}$, acts as an activation function for the input x presented. Assuming that $\mathbf{t}$ is the expected output, the associated error is calculated through $E(\mathbf{o}) = \mathbf{t} - \mathbf{o} = \mathbf{t} - f(\mathbf{iW})$. The MLP training process is given by updating the weights in $\mathbf{W}$ in order to minimize $E(\mathbf{o})$ errors.

3.4 Statistics

Statistics is a fundamental discipline to provide method and tools to find deeper insight into data. Data science and Machine Learning need the concepts of statistics to build more sophisticated models and analyses.

One of the fundamental structures of statistics is the hypothesis test which establishes a direct relationship between theory and statistics, since hypotheses can be tested with data (WEIHS; ICKSTADT, 2018).

Machine Learning uses several statistical techniques to predict and infer data. However, statistics have a greater emphasis on inference. The evaluation of a supervised classification algorithm can be performed using statistical techniques such as cross-

validation and P-values of tests (IJ, 2018).

3.5 Materials and method

3.5.1 Study Area

This project admits as a study area the region of São Paulo inserted in the hydrographic basin of the River Tamanduateí (Figure 3.1). This basin has an area of 323 km^2 and extends to the hydrographic basins of the Pinheiro, Guaió, Aricanduva and Córrego de Tapuapé rivers. This area was defined from the vicinity of a rain gauge, according to a spatial radius of 2000 m, which encompasses different flooding regions, available Tweets and a meteorological radar cell.

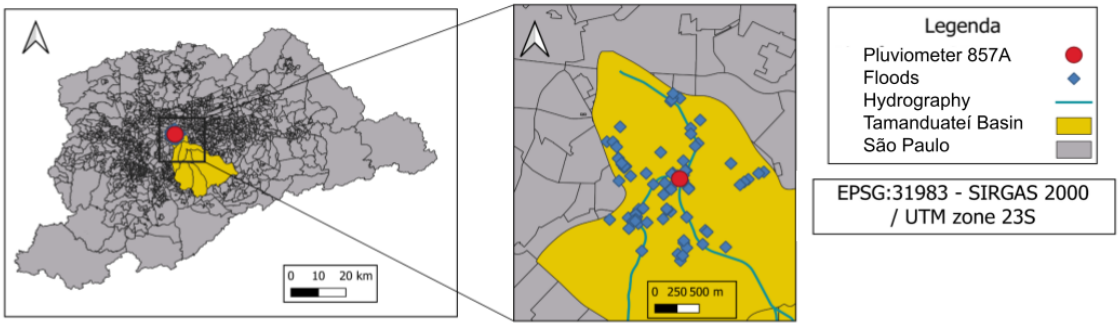

Figura 3.1 - Study Area

3.5.2 Data and tools

Twitter data were extracted through an API (*Application Programming Interface*) provided by the social network itself. The pluviometric data are collected from the pluviometer #833A, belonging to the National Center for Alerts and Natural Disasters (CEMADEN), which are made publicly available by the institution. The historical series of flooding in the study area is available through CEMADEN.

The data from the meteorological radar were extracted by a station located in the

city of São Roque. This equipment, maintained by the Department of Airspace Control (DECEA), it monitors displacement, action of clouds and instability nuclei, measuring the volume of precipitation in a given location. Furthermore, this radar has a range of 250 km, covering the entire metropolitan region of São Paulo. The radar product used is called CAPPI (*Constant Altitude Plan Position Indicator*), which has a spatial resolution of approximately 1 km and a temporal resolution of 10 minutes. For the conversion of reflectivity (dBZ) into separation rate (mm/h) the Marshall-Palmer ratio (ANDW, 1948) will be used, and then represented as "daily accumulated".

The development of the project will be guided by programming via the *Python* language. The manipulation, filtering and processing of data will be supported by the *Pandas* (VANDERPLAS, 2016) and *Numpy* (MCKINNEY, 2012) libraries.

For the application of statistical tests, the *Scipy* (VIRTANEN et al., 2020) library will be used. Similarly, the classification methods used in the research (i.e., SVM, RF and MLP) will be obtained from the Scikit-Learn library (PEDREGOSA et al., 2011).

Finally, necessary database operations will be performed with the support of the Geographic Information System *QGIS* (SAMELA et al., 2018).

3.5.3 Method

The initial design of this research proposal consists in the use of time series of floods, Tweets, data recorded by rain gauge and meteorological radar, in order to build a method for forecasting flood events. An overview of the proposal is illustrated in Figure 3.2.

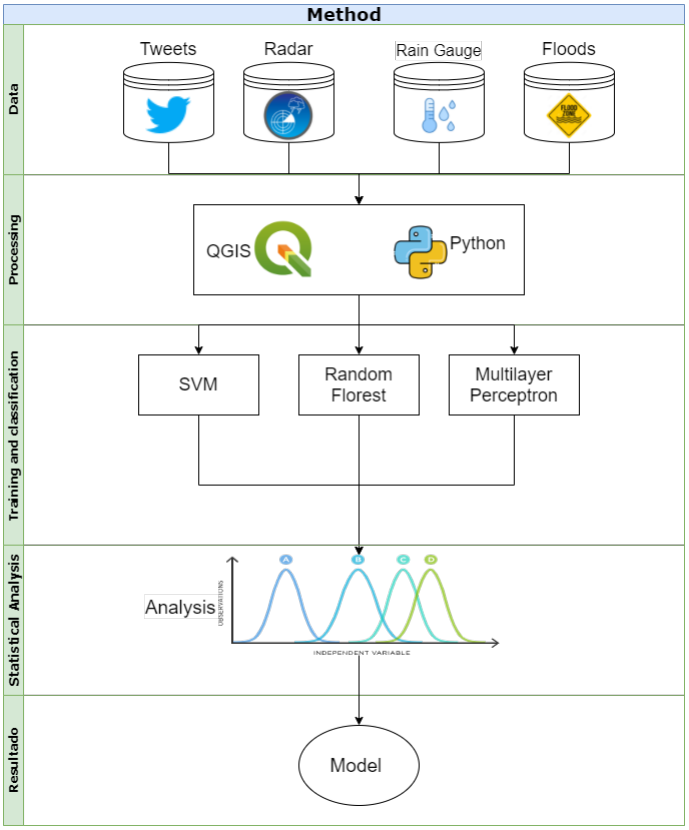

Figura 3.2 - Method

Initially, information on the number of filtered Tweets, precipitation values according to radar and rain gauge, and whether there was flooding in the analyzed days will be organized in a single file. In a second moment, this data will be submitted to the SVM, RF and MLP methods. In this phase, different subsets of attributes will be verified among the available ones. An initial division of this base will be admitted, in the proportion $\frac{2}{3} \sim \frac{1}{3}$ for the purposes of training and testing the models. The classifications to be carried out will comprise the classes “flooding” or “non-flooding”, whose accuracy will be measured by the basis of tests using measures such as kappa coefficient, F1-Score and cross validation procedure.

Subsequently, statistical tests on the significance of the results will indicate the most relevant method and attributes for building a flood warning system.

4 RESULTS AND DISCUSSION

The data were processed for the same time window and an exploratory data analysis was performed using the graph below (4.1). The pluviometric data were collected from the 833A pluviometer, in which the accumulated rain level per day was added. This rain gauge measures the rain level every 10 minutes. Radar precipitation data was collected in a cell that covers the same region as the pluviometer, thus processing the data for the analyzed temporal window. Tweets were collected within a radius of 2000m and filtered based on a list of words. The words were: 'chuva', 'chove', 'chuvoso', 'chuvosa', according to (ANDRADE et al., 2021), these words are less spatially and temporally volatile than more local and idiosyncratic terms specifically related to the city of SÃ£o Paulo (e.g. 'garoa' and 'tempestade').

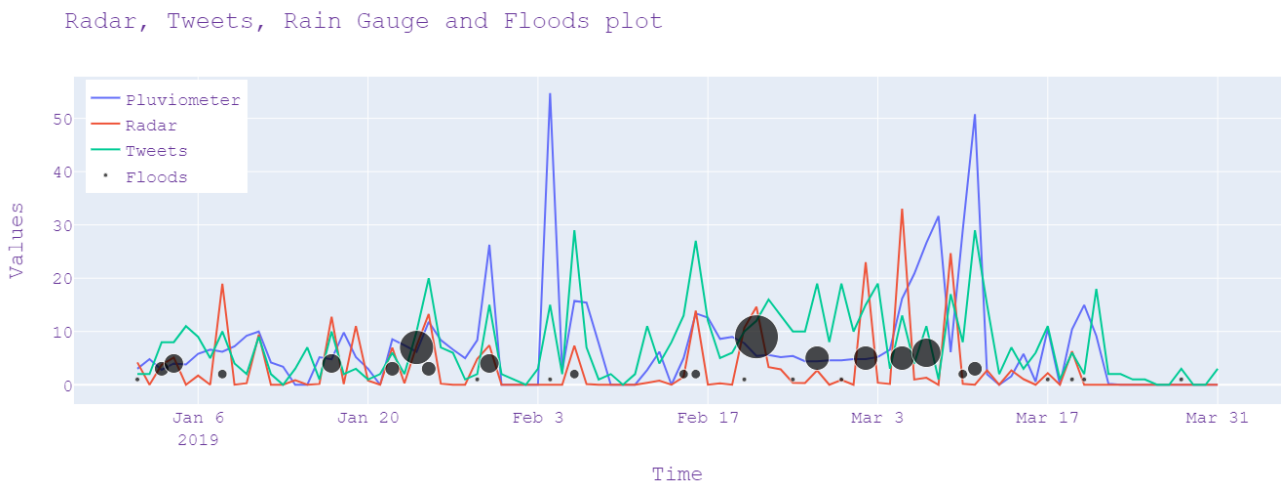

Figura 4.1 - Plot

The plotting of data indicates that on every day that flooding occurred there was an incidence of rain indicated by the rain gauge and radar and rain-related tweets. As can be seen in the figure 4.1, the frequency of flooding on a given day determines the size of the black circle, and that the days with the highest incidence of flooding

were not peaks of rain detected by meteorological equipment.

To determine the statistical relationships a series of tests were performed. First, it was necessary to determine if the attributes were a normal distribution, so that in this way parametric or non-parametric statistical tests could be applied. For this verification, the Shapiro Wilk test was used (4).

Tabela 4.1 - P-Values results

Shapiro-Wilk test $\alpha = 0.05$	
Attributes	P-value
Rain Gauge	7.175038015984347e-13
Tweet Frequence	2.0394061550632614e-07
Radar	7.194410709808908e-15
Flood	1.445436313501046e-14

In the table above 4, the p-values of all attributes were below the limit of α, discarding the null hypothesis that the distribution is normal. Based on the results, to calculate the correlations, Spearman's non-parametric test (4.2) was used for these data.

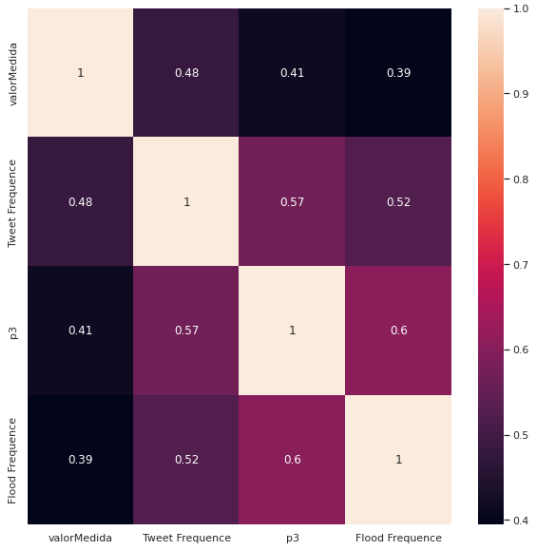

Figura 4.2 - Spearman Correlation

From the correlations, it is clear that the radar (p3), followed by the tweets, have the highest correlation with the frequency of flooding and also the most significant value. According to **??**), the cited attributes indicates a strong correlation with floods.

Furthermore, rainfall data indicate low interaction with other attributes. From the results, the most relevant correlations will also exert greater influence on the flooding prediction.

5 CONCLUSION AND PERSPECTIVES

In general, it can be observed that the value of the attributes with the floods have moderate and strong correlations. Demonstrating that there is potential to submit data to classification algorithms.

In general, it can be observed that the value of the attributes with the floods have moderate and strong correlations. In this context, it is clear that data has potential for algorithms.

In possession of primary analysis, the next step is to prepare the data for Machine Learning algorithms, optimize data quality and detect outliers.

REFERÊNCIAS BIBLIOGRÁFICAS

ALBUQUERQUE, J. P. D.; HERFORT, B.; BRENNING, A.; ZIPF, A. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. **International journal of geographical information science**, Taylor & Francis, v. 29, n. 4, p. 667–689, 2015. 2, 5

ANDRADE, S. C. de; ALBUQUERQUE, J. Porto de; RESTREPO-ESTRADA,

C.; WESTERHOLT, R.; RODRIGUEZ, C. A. M.; MENDIONDO, E. M.; DELBEM, A. C. B. 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**, Taylor & Francis, p. 1–26, 2021. 5, 11

ANDW, J. M. Mc. k. palmer,âthe distribution of raindrops with size,â. **J. Meteor**, v. 5, p. 165–166, 1948. 9

BRAGA, J. O. Alagamentos e inundações em áreas urbanas: estudo de caso na cidade de santa maria-df. 2016. 4

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001. 7

DEPARDAY, V.; GEVAERT, C. M.; MOLINARIO, G.; SODEN, R.; BALOG-WAY, S. Machine learning for disaster risk management. World Bank, 2019. 5

DOOCY, S.; DANIELS, A.; MURRAY, S.; KIRSCH, T. D. The human impact of floods: a historical review of events 1980-2009 and systematic literature review. **PLoS currents**, Public Library of Science, v. 5, 2013. 4

HANSMANN, H. Z. Descrição e caracterização das principais enchentes e alagamentos de pelotas-rs. **Universidade Federal de Pelotas, Pelotas-RS**, 2013. 4

HIRATA, E.; GIANNOTTI, M. A.; LAROCCA, A. P. C.; QUINTANILHA, J. A. Mapeamento dinâmico e colaborativo de alagamentos na cidade de são paulo. **Boletim de Ciências Geodésicas**, SciELO Brasil, v. 19, p. 602–623, 2013. 2

HORITA, F. E.; ALBUQUERQUE, J. P. de; DEGROSSI, L. C.; MENDIONDO, E. M.; UEYAMA, J. Development of a spatial decision support system for flood

risk management in brazil that combines volunteered geographic information with wireless sensor networks. **Computers & Geosciences**, Elsevier, v. 80, p. 84–94, 2015. 2, 5

HOSSAKI, C.; FEITOSA, N. et al. A twitter-based meteorological radar. **INIC**, p. 1–6, sep 2021. 5

HOSSAKI, C.; FREITAS, N. et al. Statistical relations between floods and twitter activity. **GEOINFO**, p. 1–6, oct 2021. 2

IJ, H. Statistics versus machine learning. **Nature methods**, v. 15, n. 4, p. 233, 2018. 8

LIAN, H.-C.; LU, B.-L. Multi-view gender classification using local binary patterns and support vector machines. In: SPRINGER. **International Symposium on Neural Networks**. [S.l.], 2006. p. 202–209. 6

LIU, D.; FAN, Z.; FU, Q.; LI, M.; FAIZ, M. A.; ALI, S.; LI, T.; ZHANG, L.; KHAN, M. I. Random forest regression evaluation model of regional flood disaster resilience based on the whale optimization algorithm. **Journal of Cleaner Production**, Elsevier, v. 250, p. 119468, 2020. 6

MASELLI, L. Z.; NEGRI, R. G. Integração entre estratégias multiclasses e diferentes funções kernel em máquinas de vetores suporte para classificação de imagens de sensoriamento remoto. **Revista Brasileira de Cartografia**, v. 71, n. 1, p. 149–175, 2019. 6

MCKINNEY, W. **Python for data analysis: Data wrangling with Pandas, NumPy, and IPython**. [S.l.]: "O'Reilly Media, Inc.", 2012. 9

MOSAVI, A.; OZTURK, P.; CHAU, K.-w. Flood prediction using machine learning models: Literature review. **Water**, Multidisciplinary Digital Publishing Institute, v. 10, n. 11, p. 1536, 2018. 3

NAAMAN, M. Geographic information from georeferenced social media data. **SIGSPATIAL Special**, ACM New York, NY, USA, v. 3, n. 2, p. 54–61, 2011. 4

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V. et al. Scikit-learn: Machine learning in python. **the Journal of machine Learning research**, JMLR. org, v. 12, p. 2825–2830, 2011. 9

QUEVEDO, R. P.; OLIVEIRA, G. Garcia de; GUASSELLI, L. A. Mapeamento de suscetibilidade a movimentos de massa a partir de redes neurais artificiais. **Anuario do Instituto de Geociencias**, v. 43, n. 2, 2020. 7

SAMELA, C.; ALBANO, R.; SOLE, A.; MANFREDA, S. A gis tool for cost-effective delineation of flood-prone areas. **Computers, Environment and Urban Systems**, Elsevier, v. 70, p. 43–52, 2018. 9

SANTOS, E. T. d. **Impactos econômicos de desastres naturais em megacidades: o caso dos alagamentos em São Paulo**. Tese (Doutorado) — Universidade de São Paulo, 2013. 2

SILVA, M. R. da; SANTOS, L. B. L.; SCOFIELD, G. B.; CORTIVO, F. D. 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**, v. 39, n. 1, p. 23–31, 2016. 7

SIT, M. A.; KOYLU, C.; DEMIR, I. Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: a case study of hurricane irma. **International Journal of Digital Earth**, Taylor & Francis, 2019. 2

THEODORIDIS, S.; PIKRAKIS, A.; KOUTROUMBAS, K.; CAVOURAS, D. **Introduction to pattern recognition: a matlab approach**. [S.l.]: Academic Press, 2010. 6

TINGSANCHALI, T. Urban flood disaster management. **Procedia engineering**, Elsevier, v. 32, p. 25–37, 2012. 2

VANDERPLAS, J. **Python data science handbook: Essential tools for working with data**. [S.l.]: "O'Reilly Media, Inc.", 2016. 9

VIRTANEN, P.; GOMMERS, R.; OLIPHANT, T. E.; HABERLAND, M.; REDDY, T.; COURNAPEAU, D.; BUROVSKI, E.; PETERSON, P.; WECKESSER, W.; BRIGHT, J. et al. Scipy 1.0: fundamental algorithms for scientific computing in python. **Nature methods**, Nature Publishing Group, v. 17, n. 3, p. 261–272, 2020. 9

WEIHS, C.; ICKSTADT, K. Data science: the impact of statistics. **International Journal of Data Science and Analytics**, Springer, v. 6, n. 3, p. 189–194, 2018. 7

ZHU, Z.; ZHANG, Y. Flood disaster risk assessment based on random forest algorithm. **Neural Computing and Applications**, Springer, p. 1–13, 2021. 6