UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO ACADÊMICO DO AGRESTE

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

SAULO GUILHERME RODRIGUES

MAPEAMENTO DE PERIGO DE DESLIZAMENTOS DE TERRA E INUNDAÇÕES: proposição de abordagem utilizando processamento de linguagem natural e aprendizado de máquina

Caruaru

2020

SAULO GUILHERME RODRIGUES

MAPEAMENTO DE PERIGO DE DESLIZAMENTOS DE TERRA E INUNDAÇÕES: proposição de abordagem utilizando processamento de linguagem natural e aprendizado de máquina

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Engenharia de Produção.

Área de concentração: Otimização e Gestão da Produção.

Orientador: Prof. Dr. Marcelo Hazin Alencar.

Caruaru

2020

Catalogação na fonte:
Bibliotecário – Raul César de Melo - CRB/4 - 1735

R696m Rodrigues, Saulo Guilherme.
Mapeamento de perigo de deslizamentos de terra e inundações: proposição.de abordagem utilizando processamento de linguagem natural e aprendizado de máquina / Saulo Guilherme Rodrigues. – 2020.
137 f. : il. ; 30 cm.

Orientador: Marcelo Hazin Alencar.
Dissertação (Mestrado) – Universidade Federal de Pernambuco, CAA, Programa de Pós-Graduação em Engenharia de Produção, 2020.
Inclui Referências.

1. Desastres ambientais. 2. Deslizamentos (Geologia). 3. Inundações. 4. Administração de risco. 5. Aprendizado do computador. 6. Processamento de linguagem natural (Computação). I. Alencar, Marcelo Hazin (Orientador). II. Título.

CDD 658.5 (23. ed.) UFPE (CAA 2020-043)

SAULO GUILHERME RODRIGUES

MAPEAMENTO DE PERIGO DE DESLIZAMENTOS DE TERRA E INUNDAÇÕES: proposição de abordagem utilizando processamento de linguagem natural e aprendizado de máquina

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Engenharia de Produção.

Aprovada em: 19/02/2020.

BANCA EXAMINADORA

_____

Prof. Dr. Marcelo Hazin Alencar (Orientador)
Universidade Federal de Pernambuco

_____

Prof.ª Dr.ª Maísa Mendonça Silva (Examinador Interno)
Universidade Federal de Pernambuco

_____

Prof. Dr. Rodrigo José Pires Ferreira (Examinador Externo)
Universidade Federal de Pernambuco

A minha família, em especial minha mãe Maria, meu pai Antônio e meu irmão Charles, por toda a base moral cedida.

A minha esposa Emmanuelle Carvalho pelo apoio incondicional aos meus objetivos.

AGRADECIMENTOS

Em primeiro lugar, gostaria de agradecer a minha família por todo apoio prestado desde a graduação. Minha mãe Maria Jieneilde que é a pessoa que mais me inspiro, devido a sua força de vontade e determinação para nos dar sempre o melhor. Meu pai Antônio por me mostrar como ser uma pessoa integra e me preparar para o mundo. Aos meus irmãos Charles e Chevinho por me apoiarem de todas as formas durante a graduação.

Agradeço também a minha esposa Emmanuelle por sempre acreditar em mim e apoiar meus objetivos de todas as formas. Sem você o tempo que passei cursando o mestrado seria muito mais difícil.

Agradeço aos amigos Almiro, Bruno, Dallas, e Deborah que dividiram horas de estudos comigo, cada um me ajudando de sua forma. Sem dúvidas uma das coisas mais valiosa que levarei do mestrado é nossa amizade.

Agradeço ao meu Orientador Dr. Marcelo Hazin Alencar pelo conhecimento compartilhado e pela orientação dada.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

RESUMO

Inundações e os deslizamentos de terra são desastres naturais que possuem alto potencial de dano além de apresentarem tendência crescente em termos de frequência e intensidade. Entretanto, um problema comumente enfrentado por pesquisadores ao tentarem estabelecer maneiras eficientes para gerenciar o risco de tais eventos é a indisponibilidade de dados e a dificuldade das técnicas convencionais em modelarem as relações complexas de formação de tais eventos. Tendo em vista tais questões, o objetivo do presente estudo é desenvolver um modelo para mapeamento de perigo de deslizamentos de terra e inundações com dados semiestruturados advindos de linguagem natural na forma textual, provenientes de chamadas telefônicas realizadas à órgão competente, para formar um inventário de eventos georreferenciados e com base nesse inventário mapear o perigo de deslizamentos e inundações. Para alcançar tal objetivo, foram ajustados três modelos baseados em algoritmos de aprendizado de máquina. O primeiro treinou o algoritmo *Naive Bayes* com 42000 registros textuais para classificá-los segundo seu conteúdo e formar o inventário de eventos. Os dois últimos modelos utilizaram o algoritmo *Random Forest* integrados com GIS para criar mapas de perigo de inundações e deslizamentos de terras. O modelo proposto foi testado na cidade de Recife-PE, Brasil, obtendo bom desempenho, tendo o modelo de classificação textual acurácia de 0,8671. Por sua vez o modelo para classificação do perigo de inundação obteve acuraria de 0,80 e AUC-ROC de 0,91. Por fim, o modelo de deslizamentos de terra obteve acuraria de 0,95 e AUC-ROC de 0,99. Além das avaliações quantitativas da performance, foram realizadas avaliações qualitativas, comparando os resultados gerados com notícias jornalísticas. Em todos os testes o modelo proposto apresentou resultados satisfatórios quando comparado àqueles publicados na literatura, auxiliando as partes interessadas no processo de gerenciamento de risco de desastres naturais.

Palavras-chave: Análise de perigo. Deslizamentos. Inundações. Random Forest. Naive Bayes. Processamento de Linguagem Natural. GIS. Gerenciamento de riscos.

ABSTRACT

Floods and landslides are natural disasters that have a high potential for damage in addition to an increasing trend in terms of frequency and intensity. However, a problem commonly faced by researchers when trying to establish efficient ways to manage the risk of such events is the unavailability of data and the difficulty of conventional techniques in modeling the complex relationships of the formation of such events. Given these issues, the objective of the present study is to develop an approach for mapping the danger of landslides and floods with semi-structured data from natural language in textual form, from telephone calls made to the competent body, to form an inventory of georeferenced events and based on this inventory map the danger of landslides and floods. Three models based on machine learning algorithms were adjusted. The first trained the Naive Bayes algorithm with 42,000 textual records to classify them according to their content and form the inventory of events. The last two models used the Random Forest algorithm integrated with GIS to create flood hazard maps and landslides. The proposed approach was tested in the city of Recife-PE, Brazil, obtaining good performance, with a textual classification model of 0.8671 accuracy. In turn, the flood hazard classification model obtained an accuracy of 0.80 and AUC-ROC of 0.91. Finally, the landslide model obtained an accuracy of 0.95 and AUC-ROC of 0.99. In addition to quantitative performance evaluations, qualitative evaluations were carried out, comparing the results generated with news reports. In all tests, the proposed approach showed results equal to or greater than those mentioned in the literature. In all tests, the proposed model presented satisfactory results when compared to those published in the literature, helping stakeholders in the process of managing natural disaster risk.

Keywords: Landslide hazard. Flood hazard. Random Forest. Naive Bayes. Natural Language Processing. GIS. Risk management

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE FLUXOGRAMAS

LISTA DE MAPAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

ANA	Agência Nacional de Águas
API	*Application programming interface*
AUC	Área Sobre a Curva
CCV	Côncava
CN	Curve Number
COBS	Cobertura do solo
CRVH	Curvatura horizontal
CRVV	Curvatura vertical
CVG	Convergente
CVX	Convexo
DSTEST	Distância para as rodovias
DSTRIO	Distância par o rio
DVG	Divergente
E	Leste
FFL	Formação florestal
FN	Falsos negativos
FP	Falsos positivos
GIS	*Geographic information system*
INFURB	Infraestrutura urbana
MAPTG	Mosaico de Agricultura e Pastagem
MCCV	Muito côncava
MCVG	Muito convergente
MDA	*mean decrease accuracy*
MDG	*mean decrease gini*
MDVG	Muito Divergente
NDVI	*Normalized Difference Vegetation Index*
NDWI	Normalized difference water index
NE	Nordeste
NLP	Processamento de Linguagem Natural
NVGT	Não vegetado
NW	Noroeste
OOB	Out of the Bag
PIB	Produto Interno Bruto
PLN	Planar
ROC	*Receiver Operating Characteristic*
RTL	Retilíneo
S	Sul
SE	Sudeste
SEDEC	Secretaria Executiva de Defesa Cilvil
SPI	Stream Power Index
SVM	*Support Vector Machine*
SW	Sudoeste
TF-IDF	*Term Frequency–Inverse Document Frequency*
TN	Verdaderios negativo
TP	Verdadeiros positivos
TPI	Topographic Position Index
TRI	Topographic Ruggedness Index

TWI	*Topographic wetness index*
USOS	Uso do solo
W	Oeste

SUMÁRIO

1 INTRODUÇÃO

1.1 Justificativa e relevância

Desastres são interrupções severas do funcionamento de uma comunidade ou sociedade e causa altos danos humanos, materiais, econômicos ou ambientais, as quais excede a habilidade da comunidade afetada de se reestabelecer com seus próprios recursos (United Nations Office for Disaster Risk Reduction - UNISDR, 2004). Por sua vez, desastres naturais são eventos catastróficos com origens atmosféricas, geológicas e hidrológicas que podem provocar mortes, danos a infraestrutura e interrupção do funcionamento social, nessa categoria estão, por exemplo, as inundações, deslizamentos de terras, terremotos, entre outros (XU *et al*., 2016).

Nos últimos anos houve acréscimo significante na frequência e intensidade dos desastres naturais, o que resultou em grandes perdas de vidas humanas e danos para a estrutura física (PANWAR; SEN, 2019). Ainda segundo os autores, os danos financeiros diretos são maiores para países em desenvolvimento, em termos de percentual do Produto Interno Bruto (PIB). Segundo o relatório publicados no *Intergovernmental Panel on Climate Change* - IPCC (2012), é esperado um aumento significativo das perdas e danos causados por desastres naturais devido a mudanças climáticas e o aumento da exposição e vulnerabilidade da sociedade. Mudanças climáticas podem intensificar o ciclo hidrológico, causar maiores precipitações, levando a mudanças na intensidade, frequência e severidade dos eventos desastrosos (APURV *et al*., 2015).

Assim sendo, a possibilidade de ocorrência de desastres naturais vem figurando como uma importante questão para os governantes, organizações internacionais, pesquisadores e agências de gerenciamento de emergências que objetivam obter maneiras mais eficientes para responder as potenciais consequências de tais eventos (NASCIMENTO; ALENCAR, 2016).

Dentre os desastres naturais, os deslizamentos e inundações são responsáveis por causarem diversos danos em termos financeiros, perdas de vida e interrupção da funcionalidade da sociedade. Segundo dados publicado em EM-DAT (2019), entre 1900 e 2019, os dois desastres foram responsáveis por mais de 7 milhões de mortes, mais de 3,8 bilhões de afetados e dano financeiro de quase 800 bilhões de dólares.

Devido ao potencial de dano e a tendência crescente tanto de frequência como de severidade desses tipos de desastres, vários pesquisadores têm concentrado esforços para desenvolver modelos e metodologias para o gerenciamento de risco de desastres. Segundo UNISDR (2004), gerenciamento de risco de desastres é um processo sistemático que usa decisões administrativas, organizações, habilidades e capacidades operacionais para

implementar políticas, estratégias e medidas de enfrentamento para diminuir os impactos de perigos naturais e suas consequências. Já o termo risco pode ser entendido como a probabilidade de consequências prejudiciais ou perdas resultantes da interação entre situações perigosas e condições vulneráveis naturais ou induzidos pela ação humana (UNISDR, 2004).

Para realizar análises quantitativas do perigo de deslizamentos e inundações são necessários dados históricos em quantidade e qualidade suficiente. Entretanto, como ressaltam Arabameri e outros. (2019), a falta de dados topográficos detalhados torna a tarefa de predição de inundações desafiadora. Tem e outros (2013) e ZHAO e outros (2018) afirmam que o mecanismo complexo de formação das inundações e a falta de dados são os dois principais obstáculos para o mapeamento de perigo de inundações. Su e outros (2015) discutem que frequentemente modelos de análises de deslizamentos falham devido à falta de dados sobre parâmetros importantes do modelo. Além disso, Ermini e outros (2005) relataram limitações no modelo ajustado devido falta de dados. Além das limitações impostas, a falta de dados obrigam os pesquisadores a tomarem decisões que prejudicam a precisão do modelo, como por exemplo o trabalho publicado por Guri e outros (2015), que devido a indisponibilidade de dados tiveram que considerar diferentes tipos de deslizamentos de terra como um único tipo.

Graças aos avanços recentes no poder computacional, popularização da internet e aumento da quantidade de dados compartilhados, o volume de informação disponível teve aumento significativo. Porém, tais informações necessitam passar por um processo de tratamento e interpretação, para só então, serem usadas na tomada de decisão. Tal processo, quando realizado manualmente torna-se custoso e demorado, causando uma desmotivação e consequentemente um abandono de tais dados, levando a organização a perder as informações ali contidas.

Um grau de dificuldade é adicionado ainda quando tais dados estão disponíveis como linguagem natural na forma textual. Pois para integrá-los no processo decisório é necessário compreender o significado do texto, tarefa à qual não é fácil de ser realizada devido ao tempo e quantidade de trabalho necessário. Porém, é evidente que tais dados podem conter informações úteis e que deveriam ser utilizadas para o gerenciamento de perigo de deslizamentos e inundações.

O desenvolvimento de abordagens que utilizam dados semiestruturados na forma textual para o mapeamento do perigo de desastres naturais tem como principal vantagem superar o problema de indisponibilidade de dados amplamente relatado na literatura. Além disso, ao usar algoritmos de aprendizado de máquina o processo de análise e classificação pode ser automatizado de o que proporciona maior velocidade e precisão a baixo custo.

1.2 Objetivos

1.2.1 Objetivo geral

O presente trabalho tem como objetivo geral desenvolver um modelo para mapeamento de perigo de deslizamentos e inundações que utilize dados semiestruturados advindos de linguagem natural na forma textual para formar um inventário de eventos georreferenciados e com base nesse inventário mapear o perigo de deslizamentos e inundações.

1.2.2 Objetivos específicos

Nesse contexto, o presente trabalho tem como objetivos específicos:

- Definir um padrão quantitativo para comparar o desempenho de modelos de aprendizado de máquina cujo objetivo seja o mapeamento de perigo de deslizamentos e inundações, através de uma revisão sistemática da literatura.
- Definir um processo para mapear o perigo de deslizamentos e inundações que utilize algoritmos de aprendizado de máquina, através da síntese dos resultados da revisão sistemática da literatura;
- Desenvolver um modelo para tratar e classificar de forma automática registros textuais referentes a deslizamentos e inundações, utilizando algoritmos de aprendizado de máquina e Processamento de Linguagem Natural;
- Desenvolver um modelo para calcular o grau de perigo de deslizamentos ao qual um determinado ponto está sujeito utilizando algoritmos de aprendizado de máquina integrados com Sistemas de Informações Geográficas (*Geographic information system* - GIS);
- Desenvolver um modelo para calcular o grau de perigo de inundações ao qual um determinado ponto está sujeito utilizando algoritmos de Aprendizado de Máquina integrados com GIS;
- Validar os modelos através do desenvolvimento de um estudo de caso na cidade de Recife, Pernambuco, nordeste do Brasil.

1.3 Metodologia de pesquisa

A presente seção tem como objetivo apresentar a metodologia utilizada no estudo, bem como classificá-la segundo os critérios científicos para melhor posicionamento do presente trabalho em relação as demais pesquisas.

1.3.1 Classificação da pesquisa

Em relação a natureza, a presente pesquisa trata-se de uma pesquisa aplicada. Segundo Silva e Menezes (2005), uma pesquisa aplicada tem como objetivo gerar conhecimentos para aplicações práticas e dirigidas a soluções de problemas específicos.

Em relação ao objetivo, a atual pesquisa pertence ao grupo de pesquisas descritivas, visto que segundo Gil (1991), uma pesquisa descritiva tem como objetivo descrever as características de determinada população ou fenômeno e ainda estabelecer relações entre variáveis. No presente estudo, os fenômenos estudados serão os eventos de deslizamentos e inundações e a relação entre esses eventos e variáveis escolhidas será estabelecida.

Por fim, em relação a abordagem utilizada, a atual pesquisa enquadra-se no grupo quantitativo. Segundo Turrioni e Mello (2012), a abordagem quantitativa tem como pressuposto que tudo pode ser quantificável e consequentemente utilizado para classificá-las e analisá-las. No presente estudo, tal conceito aplica-se no desejo de calcular o grau de perigo de determinada região e após isso analisar e classificar as regiões segundo tal número.

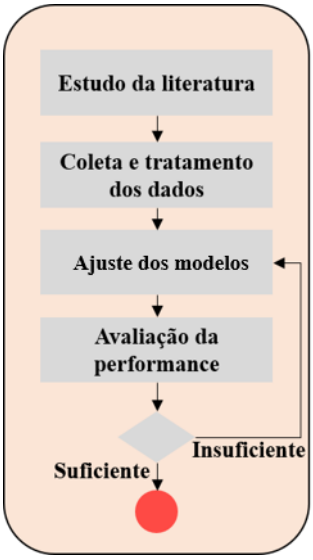
1.3.2 Processo metodológico

O processo metodológico utilizado no presente trabalho está ilustrado no Fluxograma 1. O processo foi dividido em 4 fases. Na primeira fase a literatura disponível sobre o assunto foi analisada. Na segunda fase, os dados necessários para cumprir os objetivos do trabalho foram identificados e coletados. Na terceira fase os modelos foram ajustados. Por fim, na quarta fase a performance dos modelos foram avaliadas.

A primeira fase foi responsável por estabelecer a base de conhecimento necessária sobre o assunto através de leitura técnica especializada que resultou na base conceitual disponível no capítulo 3, em seguida, foi realizada uma revisão sistemática da literatura com o objetivo de responder questões especificas sobre o tema. Toda metodologia e resultados obtidos na revisão sistemática da literatura estão expostos no capítulo 4 desse trabalho. Com a realização dessa fase foi possível identificar aspectos importantes sobre o assunto e a partir de então identificar oportunidades de melhorias nos modelos e métodos atuais.

Com a realização da análise da literatura foi possível identificar os dados necessários para ajustar o modelo, além de estabelecer padrões para comparação. Essencialmente, foram identificados dois tipos de dados. O primeiro tipo são os dados advindos de linguagem natural, no formato textual. O segundo, são dados georrefenreciados, no formato vetorial e matricial, formatos utilizados em GIS.

Fluxograma 1 – Processo metodológico utilizado no estudo



Fonte: O Autor (2019)

Com a realização da análise da literatura foi possível identificar os dados necessários para ajustar o modelo, além de estabelecer padrões para comparação. Essencialmente, foram identificados dois tipos de dados. O primeiro tipo são os dados advindos de linguagem natural, no formato textual. O segundo, são dados georrefenreciados, no formato vetorial e matricial, formatos utilizados em GIS.

Os dados na forma textual foram coletados, por meio de elaboração de uma consulta personalizada, no portal de dados abertos da cidade de Recife, disponibilizado pela Secretaria Executiva de Defesa Civil (SEDEC). Já os dados no formato vetorial e matricial foram prospectados junto às organizações governamentais competentes, como exposto no Quadro 8.

As ferramentas utilizadas para o processamento dos dados foram pacotes disponíveis na linguagem R R Core Team (2019) e a solução GIS Qgis QGIS Development Team (2019). Na linguagem R, foram utilizados pacotes para tratamento, visualização e modelagem dos dados, bem como pacotes para o ajuste dos modelos de aprendizado de máquina, descritos no capítulo 6. Já o software Qgis foi utilizado para o tratamento dos dados vetoriais e matriciais, bem como a elaboração de mapas.

Foram ajustados três modelos de aprendizado de máquina. O primeiro modelo ajustado utiliza o algoritmo *Naive Bayes* e tem como objetivo classificar os chamados advindos de linguagem natural na forma textual em quatro classes mutuamente excludentes, discutido no capítulo 5. O segundo modelo, que utilizou o algoritmo *Random Forest* para calcular o grau de perigo de deslizamento que determinado ponto está sujeito. Por fim, o terceiro modelo também

utilizou o algoritmo *Random Forest*, só que o objetivo desse modelo é calcular o grau de perigo de inundações que um determinado ponto está sujeito.

Todos os modelos foram criados utilizando algoritmos disponíveis na linguagem R (R Core Team, 2019). Entretanto, para criar os mapas de perigo foi utilizado o software Qgis R Core Team (2019). Toda a metodologia de ajuste do modelo e as discussões pertinentes estão descritas no capítulo 5 e 6 do presente trabalho.

Por fim, a avaliação da performance dos modelos foi realizada com base nas métricas identificadas na revisão sistemática da literatura. Após o cálculo dos índices de performance os resultados foram comparados qualitativamente com eventos relatados em notícias e relatórios oficiais. A discussão completa dessa fase está exposta no capítulo 6 do presente trabalho.

1.4 Estrutura do trabalho

Este trabalho está dividido em sete capítulos: introdução, contexto e descrição do problema, fundamentação teórica, revisão sistemática da literatura, modelo proposto, estudo de caso e conclusões e trabalhos futuros. Cada um dos capítulos cumpre um determinado objetivo, a saber:

- Introdução: apresentar a relevância do tema, objetivos do trabalho e metodologia geral;
- Contexto de descrição do problema: contextualizar o problema e apresentar as principais características dos eventos;
- Fundamentação teórica: apresentar e discutir os conceitos utilizados para desenvolvimento da pesquisa;
- Revisão sistemática da literatura: responder a questões especificas de pesquisa sobre o tema de mapeamento de perigo de desastres naturais com técnicas de aprendizado de máquina e estabelecer um processo para mapeamento de perigo de desastres naturais com técnicas de aprendizado de máquina, bem como estabelecer padrões de desempenho dos modelos;
- Modelo proposto: apresentar e discutir as características dos modelos propostos no presente trabalho;
- Estudo de caso: apresentar os resultados dos modelos propostos para a cidade de Recife-PE, bem como os resultados da avaliação de performance;
    - Conclusões e trabalhos futuros: apresentar o encerramento do trabalho, limitações e perspectivas para trabalhos futuros.

2 CONTEXTO E DESCRIÇÃO DO PROBLEMA DE PESQUISA

Aven (2015) argumenta que a primeira etapa que deve ser executada na etapa de análise de risco é a identificação do perigo. Essa tarefa irá definir o sucesso da análise do risco, pois caso a fonte e a magnitude do risco seja identificada de forma incorreta todas as tarefas subsequentes serão comprometidas. Para Tsai e Chen (2010) o perigo pode ser identificado através de mapas, nos quais é utilizada a distribuição de eventos passados para quantificar o perigo.

Entretanto, a indisponibilidade de dados históricos é um fator que impede a efetiva análise e quantificação do perigo de desastres naturais. Como exemplificam Stefanidis e Stathis (2013) ao discutirem que nos últimos anos, a predição do perigo de inundações tornou-se possível através da aplicação de modelos hidrológicos e hidráulicos, porém, tais modelos requerem uma grande quantidade de dados, os quais nem sempre estão disponíveis. Dessa forma, é necessário buscar novas maneiras de obter dados para viabilizar o mapeamento do perigo de desastres naturais.

Ten Veldhuis e outros (2013) propuseram um modelo de classificação de chamadas de um *call-center* municipal para análise quantitativa de risco, e utilizaram tais dados junto com a técnica análise da árvore de falha. Já o trabalho publicado por Smith e outros (2017) utilizaram dados publicados por usuários do *twitter*, uma rede social para compartilhamento de informações contendo até 240 caracteres, para monitoramento em tempo real de inundações. Além da exploração de novas fontes de dados, há também a utilização crescente de modelos de aprendizado de máquina para o mapeamento de perigo de desastres naturais (DEMIR *et al*., 2013; TEHRANY *et al*., 2015a; ZHAO *et al*., 2018). Tais modelos tem como pressuposto que os desastres que ocorrerão no futuros serão causados Segundo as mesmas causas dos desastres que ocorreram no passado (MERGHADI; ABDERRAHMANE; TIEN BUI, 2018).

As técnicas de aprendizado de máquina são importantes para superar algumas limitações existentes nos modelos determinísticos, tais como: (*i*) falta de conhecimento sobre a área de interesse, que geralmente leva a generalizações inaceitáveis; (*ii*) dificuldade na reprodutibilidade dos resultados; (*iii*) subjetividade na importância das variáveis; (*iv*) simplificações demasiadas quando os dados são incompletos; (*v*) quantidade de dados necessários muito grande nos modelos determinísticos (YILMAZ, 2010a). Além disso, os modelos físicos necessitam de grande quantidade de dados, são custosos e adequados apenas para avaliações em pequenas áreas (CHEN *et al*., 2017a; INTARAWICHIAN; DASANANDA, 2011; MERGHADI; ABDERRAHMANE; TIEN BUI, 2018). Os modelos de aprendizado de

máquina são caracterizados por uma representação mais compacta e alto potencial preditivo, utilizando poucos parâmetros e variáveis quando comparados com os modelos tradicionais (MUÑOZ *et al*., 2018).

A aquisição de informações sobre a probabilidade de ocorrência, extensão temporal e a intensidade de desastres naturais requer lidar com várias questões importantes, tais como a existência relações complexas e não lineares entre os fatores que contribuem para sua ocorrência, a falta de dados relevantes e a integração de mudanças dinâmicas que ocorrem no meio ambiente (JABBARI; BAE, 2018; POLYKRETIS; CHALKIAS; FERENTINOU, 2019). Reconhecendo os pontos fracos dos modelos tradicionais para lidar com tais questões, os pesquisadores estão mudando seus estudos para modelos de mineração de dados como *Artifical Neural Network*, *Support Vector Machine* (SVM), *Decision Tree* e *Neuro-Fuzzy Models* (POLYKRETIS; FERENTINOU; CHALKIAS, 2014).

Em estudos de desastres naturais, vários dados são necessários, mas nem sempre são fáceis de obtê-los (LEE *et al*., 2017). Apesar dos recentes desenvolvimentos, estudos que exploram novas fontes de dados e as utilizam para o mapeamento da susceptibilidade a deslizamentos de terra e inundações são poucos.

Isto posto, o presente trabalho tem como problema de pesquisa desenvolver um modelo para utilizar dados de linguagem natural semiestruturados na forma textual, advindos de registros de chamadas telefônicas, para o mapeamento de perigo de deslizamentos e inundações com algoritmos de Aprendizado de Máquina com desempenho compatível com os demais modelos disponíveis na literatura.

2.1 Conclusões do capítulo

No presente capítulo foi apresentada uma contextualização do problema enfrentado na cidade de Recife. Foram apresentadas as características dos deslizamentos e inundações, tanto do regime hidrológico bem como da urbanização dos morros, fatores que juntos aumentam a frequência de tais desastres. As responsabilidades, bem como o processo de atendimento às vítimas realizado pela SEDEC foi descrito e suas principais características comentadas. Por fim a questão de pesquisa do presente estudo foi apresentada.

3 FUNDAMENTAÇÃO TEÓRICA

3.1 Desastres Naturais

UNISDR (2004) define desastre como uma séria interrupção do funcionamento de uma comunidade ou sociedade causando perdas humanas, materiais, econômicas ou ambientais generalizadas que excedem a habilidade da comunidade ou sociedade afetada resolver com seus próprios recursos.

Desastres naturais são eventos provocados por fenômenos e desequilíbrio da natureza produzidos por fatores de origem externa independente da ação humana (BRASIL, 2009). Para a UNISDR (2004), desastres naturais consistem em um processo natural ou fenômeno que ocorre na biosfera que pode levar a eventos danosos. Tais eventos podem ser classificados pela origem, como: geológico, hidro meteorológico ou biológico.

Segundo a classificação fornecida por UNISDR (2004), os desastres são classificados em três categorias. Dentre os desastres geológicos encontram-se os terremotos, tsunamis, atividade vulcânicas, deslizamentos de terra, entre outros. Já nos desastres hidro meteorológicos encontram-se as tempestades, secas, temperaturas extremas, enchentes, entre outros. Por fim, os desastres biológicos compreendem os eventos de epidemias, contaminação devido a contato com animais ou plantas, pragas causadas por insetos, entre outros.

Goswami e outros (2018) classificam como desastres naturais terremotos, deslizamentos de terra, tempestade, enchentes. Já para De e outros (2004) os desastres naturais podem ser classificados como:

1. Desastres diretamente causados por eventos climáticos, como: furações/tufões, enchentes, secas, ondas de calor; Desastres causados indiretamente por eventos climáticos: deslizamento de terra, avalanches, incêndios florestais, fome e epidemias, etc.;
2. Desastres não relacionados com eventos climáticos, causados por eventos geofísicos: terremotos, tsunami, erupções vulcânicas, etc.

Desastres naturais podem acontecer a qualquer tempo e em qualquer lugar, tais eventos sempre causaram impactos negativos para a sociedade (HA, 2019). Segundo os dados publicados no EM-DAT (2019) sobre desastres ocorridos no período entre 1900 e 2019, os desastres naturais foram responsáveis pela maior quantidade de mortes, com mais de 32 milhões de mortes, além de causar dano financeiros maior que 8 trilhões de dólares. As enchentes ocupam o terceiro lugar em número de mortes, com mais de 6 milhões, enquanto deslizamentos de terra ocupa a sétima posição, com pelo menos 65.000 mortes.

Os desastres naturais ocorridos nos últimos anos obrigaram as autoridades públicas e organizações a aumentarem os esforços para planejar e implementar ações de gerenciamento de risco que fossem efetivas (NASCIMENTO; ALENCAR, 2016). Pois além dos danos causados aos seres humanos, como alertam Gheorghiu e outros (2014), uma das áreas mais vulneráveis e com maior concentração de ativos valiosos são as regiões industriais.

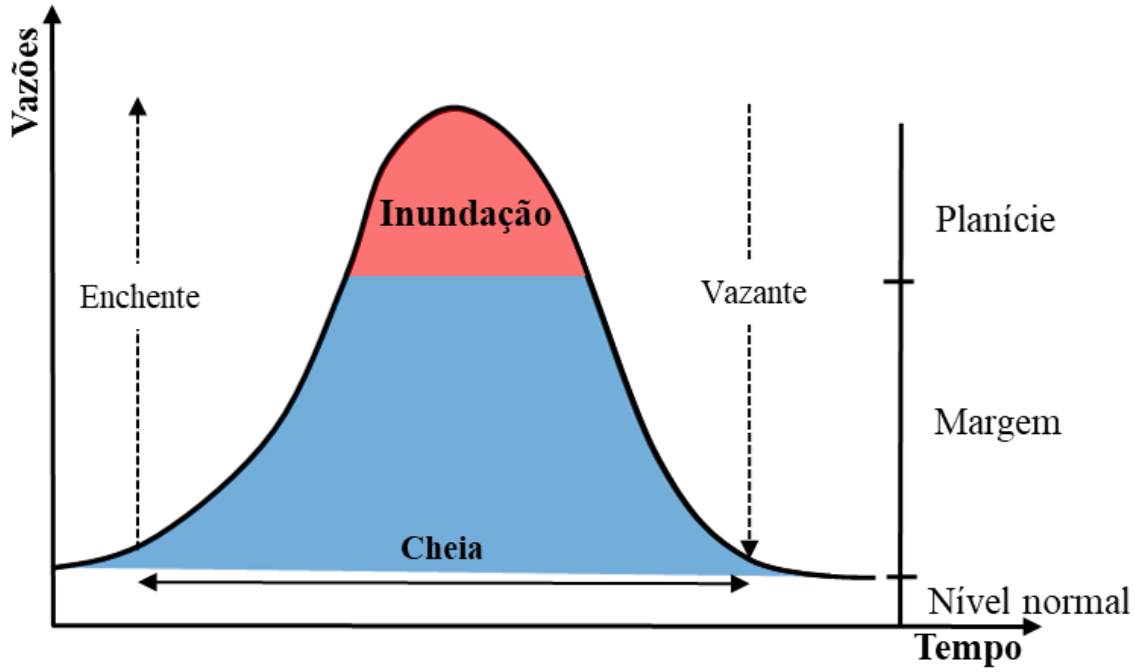
No Brasil, os desastres naturais foram responsáveis por afetar mais de 120 milhões de pessoas entre os anos de 1991 e 2012. Na Região Nordeste, mais de 55 milhões de pessoas foram atingidas, o maior número dentre todas as regiões brasileiras (BRASIL, 2013).

Nos últimos anos um crescimento significante da frequência e intensidade dos desastres naturais tem resultado em severas perdas e destruição do capital físico (PANWAR; SEN, 2019). Além disso, segundo Tehrany e outros (2015a), é esperado um aumento na ocorrência de desastres naturais, como por exemplo enchentes, devido a urbanização e desenvolvimento não planejado, aumento do desmatamento, além dos efeitos das mudanças climáticas. Youssef e outros (2011) elencam como causas do aumento da frequência de desastres naturais o desmatamento, uso intensificado do solo e o aumento populacional.

Dentre os desastres naturais, dois tipos possuem notável impacto: enchentes e deslizamentos de terra. Segundo Youssef e outros (2011), enchentes podem influenciar vários aspectos da vida humana devido aos seus efeitos destrutivos e alto custo para mitigação. Tehrany e outros (2015a) afirmam que entre os vários tipos de desastres naturais, as enchentes são consideradas como um dos mais devastadores. Já para Pradhan (2010), deslizamentos de terras são os maiores desastres naturais, entre os desastres geológicos, e em cada ano é responsável por danos enormes envolvendo tanto custos diretos como indiretos.

Antes de mais nada, é importante distinguir os termos cheia, enchente, inundação, enxurrada e alagamentos. Cheia é o período do ano hidrológico associado à ocorrência das maiores precipitações. Enchentes é uma elevação do nível de água de um rio, acima de sua vazão normal, caracterizando a ascensão do hidrograma no período de cheia. Inundação é um extravasamento da vazão do rio para fora de sua calha secundária, ocupando a planície de inundação e ocorre durante a enchente. Enxurrada é uma inundação brusca, que ocorre em terrenos de alta declividade. Por fim, alagamento é o acúmulo de água em áreas urbanas por falha do sistema de drenagem (MIGUEZ; GREGORIO; VERÓL, 2018). Uma representação visual desses conceitos é mostrada no hidrograma explicitado na Figura 1. Para fins de simplificação, no presente trabalho o termo inundação será utilizado para descrever tanto inundações como alagamentos.

Figura 1 – Definição de cheia, enchente e inundação



Fonte: adaptado de Miguez, Gregorio E Veról (2018)

Em diversas situações, geólogos, engenheiros e outros profissionais utilizam definições ligeiramente diferente sobre deslizamentos. Para fins desse estudo, será utilizada a definição dada por Highland e Bobrowsky (2008), que considera como deslizamentos o movimento de descida do solo, de rochas e material orgânico, sob o efeito da gravidade, bem como a formação geológica resultante de tal movimento.

3.2 Gerenciamento de riscos em desastres naturais

A sociedade lida com riscos todos os dias, tanto que analisá-los tornou-se uma atividade comum aos seres humanos. Nas atividades diárias o risco também está presente, por exemplo: ao caminhar na rua, usar transporte público para ir ao trabalho ou até mesmo ao ingerir comidas gordurosas. Apesar da sociedade lidar com o risco diariamente, ainda não há um consenso total a respeito da definição do termo risco na literatura (ALMEIDA *et al*., 2015).

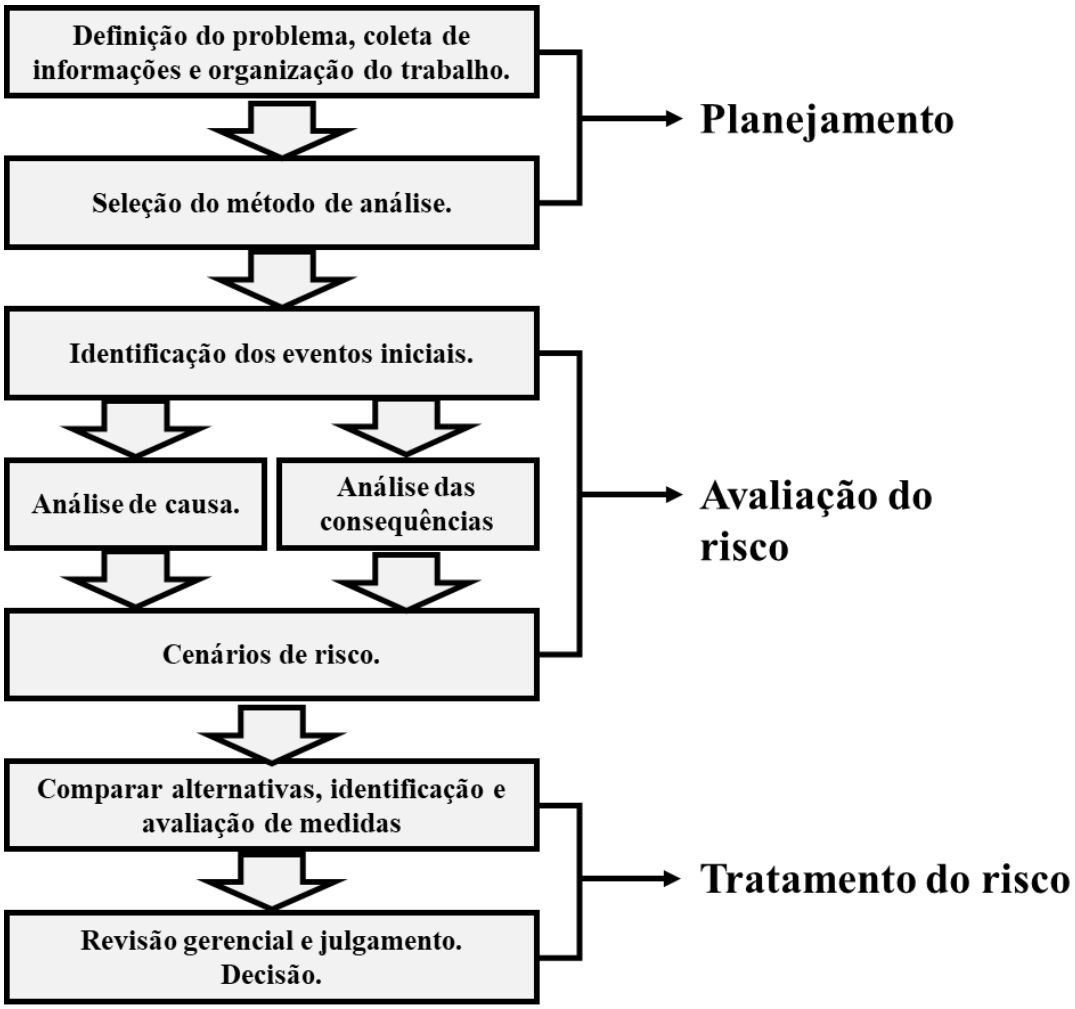
Aven (2015) define risco como sendo a união de um evento e suas consequências. Não sendo necessário ter certeza sobre a ocorrência do evento, muito menos das suas consequências. Essa incerteza pode ser representada através de probabilidades, baseada em conhecimento

prévio sobre o evento. Para Aven e Reniers (2013) a interpretação do que realmente significa tal probabilidade pode afetar fortemente o processo de decisão.

Garrick e outros (2001) definem risco como uma função do cenário de risco, da probabilidade do cenário de risco acontecer e do vetor de dano. Já Campbell (2005) afirma que uma das definições básicas para risco é junção de uma ação e da respectiva consequência.

Aven (2015) propõe um processo de análise de risco dividido em 3 fases, como mostra o Fluxograma 2. A primeira fase consiste no planejamento. A segunda fase é responsável por avaliar o risco. Por fim, na terceira fase ocorre o tratamento do risco.

Fluxograma 2 – Processo de análise de risco



Fonte: Adaptado de Aven (2015)

Ainda segundo esse autor, os métodos de análise de risco se dividem em três categorias principais, segundo mostra o Quadro 1.

Quadro 1 – Categorias dos métodos de análise do risco

Categoria principal	Tipo da análise	Descrição
Análise simplificada do risco	Qualitativa	Estabelece o cenário de risco usando brainstorming e discussões em grupo.
Análise padronizada do risco	Qualitativa ou quantitativa	Análise de risco mais formalizada, na qual utiliza metodologias como HAZOP para estabelecer o cenário de risco.
Análise baseada em modelo	Predominantemente quantitativa	Usa técnicas como análise de árvore de falha, análise de árvore de eventos de falha para calcular o risco.

Fonte: Adaptado de Aven (2015)

Tanjin Amin e outros (2019) dividiram os métodos de análise de risco em 3 categorias segundo a frequência de utilização. Altamente frequente: Análise de árvore de Falha (*Fault Tree Analysis* - FTA), Estudo de perigo e operacionalidade (*Hazard and Operability Studies* - HAZOP) e Teoria *Fuzzy*. Frequência moderada: Rede Bayesiana (*Bayesian Network* – BN), Análise Hierárquica do Processo (*Analytic Hierarchy Process*- AHP), Árvore de Eventos de Falha (*Event Tree Analysis* – ETA). Pouco frequente: Análise das Camadas de Proteção (*Layers of Protection Analysis* – LOPA), Rede de Petri (*Petri-net* – PN), e Análise *Bow-Tie*.

Além dos métodos tradicionais de análise de risco, vários estudos foram realizados explorando novos modelos. Por exemplo, Wang e outros (2015b) propuseram um modelo baseado no algoritmo *Random Forest* para calcular o índice de risco de cada região. Um dos principais argumentos dos autores foi a falta de linearidade na relação entre os índices e as ocorrências das inundações. Yao e outros (2008) ajustaram um modelo baseado no método SVM para o zoneamento de perigo de deslizamentos de terra e obtiveram resultados com acurácia de 0,9039.

Um conceito importante para o presente trabalho é o de Risco de Desastres. IPCC (2012) define tal conceito como a probabilidade, em um período específico, de alterações severas no funcionamento normal de uma comunidade ou sociedade devido a eventos físicos perigosos que interagem com condições sociais vulneráveis, levando a largas perdas humanas, materiais, econômicas, ou ambientais, de modo que seus efeitos necessitem de respostas imediatas para satisfazer as necessidades humanas e que requeiram suporte externo para a total recuperação.

Por sua vez, o gerenciamento de risco de desastres naturais pode ser entendido como um processo sistemático de uso de decisões administrativas, organizações, habilidades e capacidades operacionais para implementar políticas e estratégias, bem como capacidade de

lidar com a comunidade e sociedade para diminuir os impactos dos perigos naturais e desastres ambientais e tecnológicos relacionados. Isso inclui todas as formas de atividades, podendo ser medidas estruturais e não estruturais para evitar (prevenção) ou limitar (mitigação) os efeitos adversos dos perigos.

Para Tsai e Chen (2010) o processo de gerenciamento de risco de desastres pode ser representado por duas características fundamentais: formação do risco de desastre e estratégias de gerenciamento.

Para que o risco seja formado é necessário a presença de três elementos: a fonte do perigo, a vulnerabilidade, e a exposição da comunidade. O perigo pode ser entendido como um evento físico, fenômeno ou atividade humana potencialmente danosa que pode causar danos a vida, sociais, financeiros ou ambientais (UNISDR, 2004). Segundo Papathoma-Köhle e outros (2017), a definição clássica de vulnerabilidade representa o grau de perda de um dado elemento, ou conjunto de elementos, dentro de uma área afetada por um desastre, tal vulnerabilidade é expressa entre 0 (sem danos) e 1 (dano total). Por fim, a exposição representa o grau de presença de pessoas, meios de sobrevivência, recursos ambientais, infraestrutura, ativos sociais, econômicos e culturais em locais que podem ser afetados negativamente por eventos físicos e portanto, estão sujeitos a possíveis danos ou perdas futuras (IPCC, 2012).

As estratégias de gerenciamento que podem ser adotadas por um gestor incluem: Retenção, mitigação e transferência. Retenção do risco refere-se a ação de assumir e aceitar a existência do risco, sem transferência de responsabilidade para terceiros (TSAI; CHEN, 2010). A mitigação do risco, segundo IPCC (2012), são ações tomadas que visam limitar condições futuras adversas uma vez que o desastre está materializado, como explicam Miguez e outros (2018), diferentemente da etapa de prevenção, nas ações de mitigação já existem elementos expostos e é necessário atuar nos componentes do risco para evitar seus efeitos danosos. Por fim, a transferência do risco é o processo de mudança formal ou informal das consequências de um risco particular de uma parte para a outra, na qual a parte que assumirá o risco será beneficiada de alguma maneira (IPCC, 2012).

Apesar da dificuldade em prever os desastres naturais, a avaliação da vulnerabilidade, mitigação do risco e planos de gerenciamento de emergência podem reduzir os impactos dos eventos e facilitar a recuperação da área atingida (FRIGERIO *et al*., 2016).

A problemática de gestão de riscos de desastres hidrológicos é abordada de forma multidisciplinar e em várias áreas de estudo. Por exemplo: Alexander e outros (2016) elaboraram um *framework* para avaliar as políticas governamentais para o tratamento do risco

de enchentes. Outro exemplo: Alves e outros (2018) adotaram uma abordagem multicritério para seleção de infraestrutura sustentável para reduzir o risco de inundações.

Almeida e outros (2015) chamam atenção para o fato que em muitas situações reais, as consequências do risco são multidimensionais, considerando dimensões como financeiro, confiabilidade e segurança. No contexto de riscos desastres naturais, tal característica foi incorporada a partir de trabalho como Alves e outros (2018) e Kubal e outros (2009).

3.3 Processamento de linguagem natural

Na atual sociedade orientada por informações, o número de documentos disponíveis de forma *online* vem crescendo exponencialmente. Tais documentos armazenam informações que podem ser utilizadas pelos tomadores de decisões em suas organizações. Porém, fazer isso manualmente requer muito esforço e tempo, devido ao grande volume de informações que devem ser analisadas e tratadas. Consequentemente, empregar sistemas inteligentes para detectar informações essenciais automaticamente de documentos textuais garante que as organizações tomem as decisões certas, no tempo certo e que proporcionam vantagens competitivas (HADI, 2013).

A área de estudo desse tema é conhecida como Processamento de Linguagem Natural (*Natural language Processing* – NLP), que descreve uma família de algoritmos que podem capturar informações chaves de textos livres e converter em um formato estruturado para análise. Esses algoritmos podem usar métodos estatísticos e linguísticos para capturar informações sobre variação das palavras, ambiguidades, inconsistências em documentos ou ainda outras informações de interesse (MAGANTI *et al*., 2019). NLP teve início na década de 1950 como o resultado da interação entre inteligência artificial e linguística (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).

Tarefas que podem parecer fáceis para os humanos representam desafios enormes para as máquinas. Imagine, por exemplo, que um engenheiro escreveu a simples frase: “O desgaste no eixo foi causado por falta de lubrificação”. Qualquer ser humano que saiba ler irá facilmente compreender o significado dessa frase, porém, para as máquinas, não é uma tarefa fácil, visto que essas não possuem a capacidade de compreensão simbólica humana, muito menos a capacidade de formulação de conceitos a partir de poucos dados de entrada. Para tornar as máquinas capazes de entender o significado da frase acima é necessário utilizar técnicas estatísticas sofisticadas.

Felizmente, segundo Moreno (2019) nos últimos 60 anos, a evolução do NLP possibilitou a comunicação entre humanos e computadores através de interfaces de conversação. No início

1980, a maioria das abordagens de NLP eram baseadas em conjuntos de regras complexas e artesanais, tal abordagem provocava ambiguidade nas análises realizadas, tornando possível várias interpretações para uma única sequência de palavras (HAN; KWOH, 2019). Com a aplicação dos algoritmos de Aprendizado de Máquina, tal característica foi contornada, porém, como alertam Han e Kwoh (2019), as abordagens estatísticas precisam ser treinadas e são dependentes do contexto do problema.

Segundo Nadkarni e outros (2011), as abordagens tradicionais (*hand-written*) levam a dois problemas quando lidam com grandes volumes de linguagem natural. O primeiro problema citado é que existem diversas relações linguísticas necessárias para extrair o significado de uma sentença, o que torna as regras muito grandes e imprevisíveis, dando espaço para interpretações ambíguas. Por fim, um outro problema destacado é que as regras escritas são feitas baseadas na gramática formal, porém, nas situações práticas frequentemente a linguagem não segue tal padrão, tornando as regras não confiáveis.

Todos esses fatos corroboraram, na década de 1980, para o surgimento de melhorias no NLP. Como sumarizado por Nadkarni e outros (2011):

- Aproximações simples e robustas substituíram análises profundas;
- A etapa de avaliação tornou-se mais rigorosa;
- A utilização de métodos de aprendizado de máquina tornou-se proeminente;
- Grandes volumes de texto foram utilizados para treinar os algoritmos de aprendizado de máquina, fornecendo padrões para comparação.

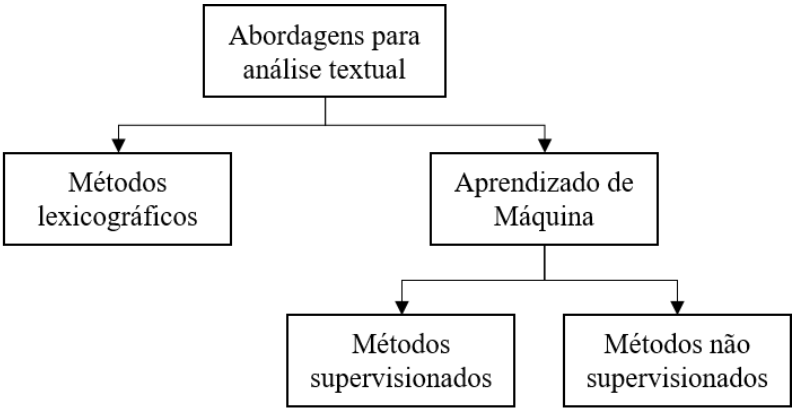
Dentro do escopo de NLP, classificação textual é uma área especialmente importante para o presente estudo e amplamente utilizada. A classificação textual tem como foco classificar documentos em uma ou mais classes predefinidas tomando como base seu conteúdo (RADAIDEH; KHATEEB, 2015). Ainda segundo os autores, a área recebeu muita atenção nos últimos anos, onde vários métodos foram usados tais como: SVM; *Naive Bayes*; *k-Nearest Neighbour*, *Artificial Neural Networks*; entre outros.

Segundo Hadi e outros (2018), construir sistemas de classificação de texto automatizados é um dos mais importantes tópicos nas áreas de mineração de dados e aprendizado de máquina, pois classificações manuais requerem alta acurácia e consomem muito tempo, enquanto sistemas automatizados permitem que o processo seja mais eficiente e rápido.

Hartmann e outros (2019) dividem as abordagens para análise automática de texto em abordagens lexicográficas e abordagens baseadas em aprendizado de máquina. As abordagens lexicográficas são aquelas que utilizam o repertório lexical e valor semântico das palavras para

classificar um determinado conteúdo. Já as abordagens baseadas em aprendizado de máquina utilizam técnicas estatísticas para classificar o documento com base em seu conteúdo. Na Figura 2 estão expostas as abordagens e as tarefas para as quais cada abordagem é indicada, segundo Hartmann e outros (2019).

Figura 2 – Abordagens para análise automática de texto



Tarefas			
Classificação de texto			
• Sentimento	Sim	Sim	Não
• Conteúdo	Não	Sim	Não
Modelagem de tópico	Não	Não	Sim

Fonte: adaptado de (HARTMANN et al., 2019)

Antes de efetivamente ajustar um modelo estatístico para os documentos é necessário prepará-los para tal. Radaideh e Khateeb (2015) explicam que o principal objetivo da fase de pré-processamento é representar cada documento como um vetor de termos, limpar o texto e reduzir o tamanho do vetor. Para os autores, as principais etapas da fase são: criação dos *tokens*; remoção de *stop words*; estemização e; seleção dos termos importantes.

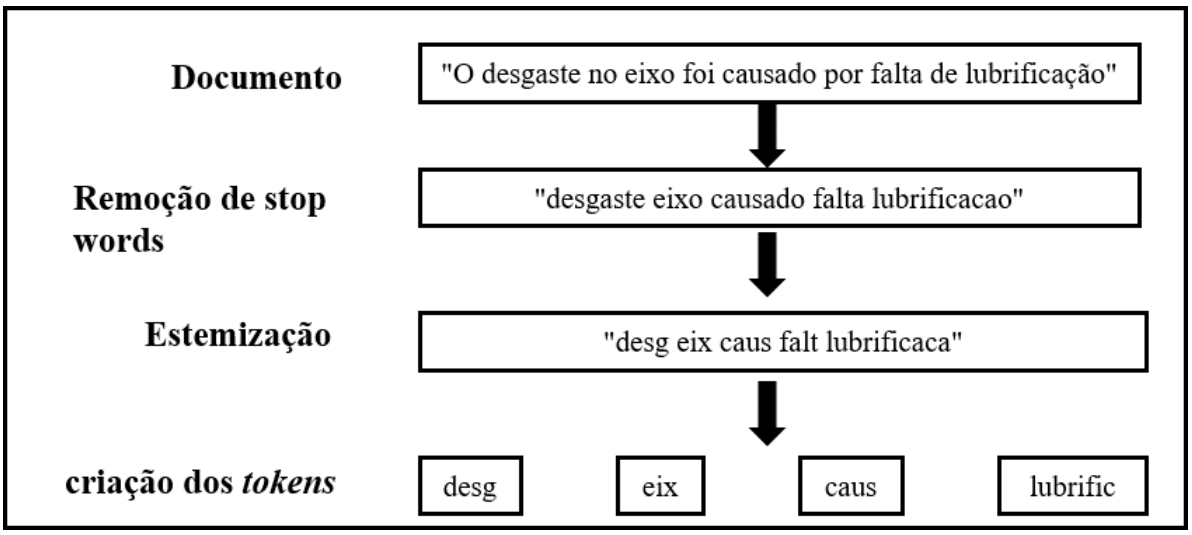
A criação dos *tokens* refere-se a tarefa de dado um documento com *m* palavras, esse documento será representado por um vetor de termos com tamanho *m*. Cada elemento do vetor representa um *token*. Nessa etapa também são retiradas todas as acentuações gráficas.

Na etapa de remoção das *stop words* são retiradas todas as palavras que não possuem valor semântico, ou seja, que não agregam valor a frase. Por exemplo, na frase: “O desgaste no eixo foi causado por falta de lubrificação”. São consideradas *stop words* as palavras: o, no, foi, por e d. Perceba que caso essas palavras sejam removidas da frase o sentido ainda pode ser capturado, porém, se deixar somente as *stop words* na frase, nenhum sentido pode ser extraído

da sentença. Um dos principais benefícios dessa fase é diminuição de palavras para serem analisadas.

A estemização é a tarefa de reduzir as palavras em seu tronco ou raiz. Por exemplo, as palavras deslizamentos, deslizou, inundação e inundou, podem ser representadas pelos troncos “desliz” e “inund”. Tal tarefa faz com que palavras com o mesmo valor semântico sejam representadas pelos mesmos caracteres, reduzindo a quantidade de palavras a serem analisadas. Vale ressaltar que o tronco não precisa ser igual a raiz morfológica da palavra, isso irá variar a depender do algoritmo utilizado. A Figura 3 exibe um exemplo para as três etapas descritas até o momento.

Figura 3 – Pré-processamento de texto



Fonte: O Autor (2019)

Por fim, para cada termo é atribuído um peso que servirá para filtrar os termos mais importantes, com objetivo de reduzir o tamanho do vetor e selecionar apenas os que possuírem mais valor para a análise. Uma abordagem frequente é utilizar a Frequência de Termos pela Frequência Inversa do Documento (*Term Frequency–Inverse Document Frequency* - TF-IDF), A Frequência de Termos (TF) determina a quantidade de vezes que o termo *i* apareceu no documento *j*. Por sua vez, o Frequência Inversa do Documento (IDF) é definida pela Equação (3.1).

$$IDF = log\left(\frac{N}{n_i}\right) \tag{3.1}$$

Onde N representa o número total de documentos analisados e n_i é o número de documentos que contém o termo i, dentre todos os documentos analisados.

Após o cálculo da frequência de cada termo e a frequência inversa de cada documento para cada termo, o peso é obtido segundo a Equação (3.2). A abordagem do TF-IDF é baseada no pressuposto que os termos mais significantes no documento são os que aparecem menos vezes na coleção de documentos (SABBAH *et al*., 2017). Por exemplo, se em uma coleção as palavras "solicitou", "falou", "relata" e "urgência" aparecem em todos os documentos, logo estas palavras não possuem valor preditivo para caracterizar o documento, pois aparecem mais ou menos na mesma quantidade e cada um dos documentos. Porém, as palavras "deslizamentos" e "deslizou" aparecem em poucos documentos, dessa forma, tais palavras são mais significativas para classificar estes documentos.

$$TF - IDF = TF \times IDF \tag{3.2}$$

O TF-IDF faz uma ponderação entre quantas vezes o termo apareceu com a quantidade de documentos que possui o termo.

3.4 Aprendizado de máquina

Uma das primeiras definições de Aprendizado de Máquina foi publicada por Samuel (1959). Segundo o autor, aprendizado de máquina é o processo de ensinar máquinas através da experiência. Shalev-Shwartz e Ben-David (2013) apresentaram um conceito mais abrangente. Para os autores, aprendizado de máquina tem como objetivo programar computadores de modo que eles aprendam através de entradas disponibilizadas para os mesmos. Ainda segundo os autores, o aprendizado é o processo de converter experiência em expertise ou conhecimento. A entrada para os algoritmos de aprendizado são os dados de treino, representando a experiência, e a saída do modelo representa algum tipo de informação que trará conhecimento.

Para Jordan e Mitchell (2015), um problema de aprendizado pode ser definido como o problema de melhorar alguma medida de performance executando alguma tarefa, com algum tipo de treinamento através experiências. No caso do presente estudo, a tarefa pode ser representada por atribuir o grau de perigo de desastre para uma determinada localidade. A métrica de performance pode ser a acurácia da previsão. Por fim, a experiência pode ser representada pelo conhecimento prévio da localidade de ocorrência dos desastres.

Para Jordan e Mitchell (2015), o aumento da disponibilidade de grandes volumes de dados em todas as áreas do conhecimento humano provocou uma demanda pelo entendimento de

algoritmos de aprendizado de máquina. Ainda segundo os autores, a área de estudo desse tópico situa-se entre ciência da computação, estatística e uma variedade de outras disciplinas concentradas em melhoria automática ao longo do tempo e tomada de decisão em ambientes de incerteza.

Os algoritmos de aprendizado de máquina podem ser classificados segundo o tipo de aprendizagem. Aprendizado supervisionado utiliza dados históricos com variáveis reposta previamente identificadas. No aprendizado não supervisionado os dados não possuem variável resposta, o aprendizado ocorre com base em medidas de similaridade ou distância entre as observações. Uma abordagem híbrida é o aprendizado semi-supervisionado, o qual existe variável resposta identificadas para apenas uma parte do conjunto de dados. Por fim, o aprendizado por reforço, faz uso apenas de dados do estágio inicial e aprende continuamente através da interação com o ambiente no qual está inserido, em que carros autônomos e robôs são exemplos de tecnologias que utilizam tal tipo de aprendizado (RAMASUBRAMANIAN; SINGH, 2017).

O processo de aprendizado de máquina pode ser dividido em 5 etapas (LANTZ, 2015). A primeira etapa é a coleta de dados. Segunda etapa é a exploração e preparação dos dados coletados. A terceira será a etapa onde o modelo será treinado. Após o treino, na quarta etapa, a performance do modelo será avaliada. Por fim, na quinta etapa, melhorias no modelo podem ser implementadas segundo as necessidades especificas de cada problema.

Atualmente existe uma grande disponibilidade de algoritmos de aprendizado de máquina, dos quais serão abordados apenas dois deles no presente trabalho. O primeiro algoritmo a ser abordado será o *Naive Bayes*, método de aprendizado estatístico baseado no Teorema de Bayes. O segundo, o algoritmo *Random Forest*, derivado de árvores de decisão. Maiores detalhes sobre os motivos pelos quais tais algoritmos foram escolhidos serão apresentados em seções subsequentes.

3.4.1 Naive Bayes

O *Naive Bayes* é um algoritmo de aprendizagem estatística baseado no Teorema de Bayes. Seus principais usos estão relacionados a classificação de texto, classificação de documentos e filtro de spam (RAMASUBRAMANIAN; SINGH, 2017) .

Antes de mais nada, é importante entender o funcionamento do Teorema de Bayes. Tal teorema define a probabilidade condicional entre dois eventos, como exposto na Equação (3.3).

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{3.3}$$

Onde:

P(A|B) - Probabilidade a posteriori do evento A acontecer dado que o evento B aconteceu.

P(B|A) - Probabilidade a priori do evento B acontecer dado que o evento A aconteceu.

P(A) - Probabilidade do evento A acontecer

P(B) - Probabilidade do evento B acontecer

Um teorema fundamental no entendimento do tópico é o Teorema da Probabilidade Total, que considerando que o espaço amostral pode ser dividido em *n* eventos mutuamente exclusivos A_i, i = 1, 2 ... n, permite reescrever o Teorema de Bayes na forma estendida, como mostra a Equação (3.4).

$$\mathrm{P(A_i|B)} = \frac{\mathrm{P(B|A_i)} \cdot \mathrm{P(A_i)}}{\mathrm{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_n)P(A_n)}} \tag{3.4}$$

Dado que tenha um banco de dados com *n* variáveis independentes, representada pelo vetor x = (x_1, ..., x_n), então uma observação pode ser classificada com a probabilidade P(C_k| x_1| ... | x_n) em qualquer uma das K classes C_k (RAMASUBRAMANIAN; SINGH, 2017). Tal probabilidade pode ser expressa pela Equação (3.5).

$$P(C_k|x) = \frac{P(C_k) \cdot P(x|C_k)}{P(x)} \tag{3.5}$$

A Equação (3.5) pode ser reescrita utilizando a regra da cadeia, a qual toma a forma apresentada na Equação (3.6).

$$P(C_k|x) = \frac{P(C_k) \cdot \ P(x_1|C_k) \cdot P(x_2|C_k) \cdot P(x_3|C_k) \ldots \ P(x_n|C_k)}{P(x_1) \cdot P(x_1) \cdot P(x_1) \cdot \ldots \ P(x_n)} \tag{3.6}$$

Em problemas de classificação de texto, o vetor de variáveis *x* é representado pela frequência de cada palavra. Então, dessa forma, para saber se uma dada observação pertence uma classe C_k, é necessário calcular a probabilidade condicional da classe C_k dado um determinado vetor de frequência de palavras x = (x_1, ..., x_n). Calcula-se a probabilidade para cada uma das *k* classes, a observação será classificada para a classe *k* com maior probabilidade (ARTISSA; ASROR; FARABY, 2019).

Segundo Lantz (2015) o nome "*naive*", que significa ingênuo, vem do fato que o algoritmo pressupõe que as variáveis sejam independentes e igualmente importantes para o problema. Apesar disso, argumenta o autor, *Naive Bayes* consegue altas performances mesmo quando tais requisitos não são totalmente atendidos, alcançando a mesma ou até melhor performance que algoritmos mais sofisticados.

Lantz (2015) afirma ainda que tal algoritmo possui vantagens como: simplicidade, velocidade e eficiência; lida bem com ruídos e valores omissos; requer poucos dados para treino e consegue lidar com grandes volumes de dados com a mesma eficiência; facilidade para obter a probabilidade estimada para a predição. Porém, entre os defeitos estão: possuir pressupostos de independência e igualdade de importância; não lida bem com banco de dados com grandes números de variáveis numéricas e; estimar as probabilidades é menos confiável que prever as classes.

P. Domingos e Pazzani (1997) publicaram um importante trabalho a respeito da confiança do algoritmo *Naive Bayes* em situações com variáveis com alta dependência entre si. Os autores mostram que o algoritmo consegue desempenhar bem a tarefa de classificação mesmo em situações nas quais as variáveis possuem alto índice de dependência. Além disso, foi mostrado empiricamente que não é necessária independência entre os atributos para ser ótimo sob função de perda zero-um. Por fim, os autores evidenciaram através de experimentos que mesmo nas situações em que o algoritmo não é ótimo, consegue performance melhor que vários outros algoritmos mais complexos.

Como evidência que o Algoritmo *Naive Bayes* consegue bom desempenho em tarefas de classificação de texto, vários trabalhos foram publicados. Frank e Bouckaert (2006) utilizaram o algoritmo para classificação de texto com banco de dado desbalanceado, mostrando que o modelo conseguiu atingir acurácia superior a 0,9. Pranckevičius e Marcinkevičius (2017) compararam cinco modelos de aprendizado de máquina para a classificação de textos e chegaram à conclusão que o *Naive Bayes* obteve a melhor performance. Häberle e outros (2019) publicaram um estudo no qual utilizam, entre outros, o algoritmo *Naive Bayes* para classificar prédio em comerciais ou residenciais, utilizando informações textuais publicadas no *twitter*. Hartmann e outros (2019) compararam algoritmos de aprendizado e de máquina e abordagens lexicográficas. Dentre os resultados foi evidenciado empiricamente que o algoritmo *Naive Bayes* apresentou melhor desempenho entre os testados. Além do algoritmo *Naive Bayes*, o *Random Forest* apresentou bons resultados para a tarefa de classificação textual.

3.4.2 Random Forest

Entende-se por *Random Forest* um classificador composto por uma coleção de classificadores estruturados baseados em árvores de decisão $\{h(\boldsymbol{X}, \Theta_k), k = 1, \ldots\}$ onde $\{\Theta_k\}$ são vetores aleatórios identicamente e independentemente distribuídos e cada árvore contribui com uma unidade de voto para a classe mais popular do vetor de entrada *X* (BREIMAN, 2001). A predição do *Random Forest* é obtida através de voto majoritário (BEN-DAVID; SHALEV-SHWARTZ, 2014). O Quadro 2 mostra o algoritmo *Random Forest*.

Quadro 2– Algoritmo Random Forest

Algoritmo *Random Forest para Regressão e classificação*

var

Inteiro: B, N, m, n_{min};

Inicio

Para b = 1 até B, faça

Crie a amostra aleatória Z^* de tamanho N dos dados de treino com reposição;

Crie a árvore T_b com a amostra Z^*, através da repetição dos passos abaixo para cada nó final da arvore, até que o tamanho mínimo do nó n_{min} seja alcançado:

i. Selecione *m* variáveis aleatoriamente das M variáveis disponíveis;
ii. Escolha a melhor candidata para o corte das *m* variáveis;
iii. Divida o nó em dois nós filhos.

Fim para

Retorne o conjunto de árvores $\{\boldsymbol{T_b}\}_1^B$

Fim

Para realizar as predições:

Regressão: $\boldsymbol{f_{rf}^B(x)} = \frac{1}{B}\sum_{b=1}^{B} \boldsymbol{T_b(x)}$

Classificação: Seja $\widehat{\boldsymbol{C}}_b(\boldsymbol{x})$ a classe predita pela b-enésima árvore da floresta aleatória. Então, $\widehat{\boldsymbol{C}}_{\boldsymbol{rf}}^{\boldsymbol{B}}(\boldsymbol{x}) = \boldsymbol{voto\ majorit}$**ário** $\{\widehat{\boldsymbol{C}}_b(\boldsymbol{x})\}_1^B$

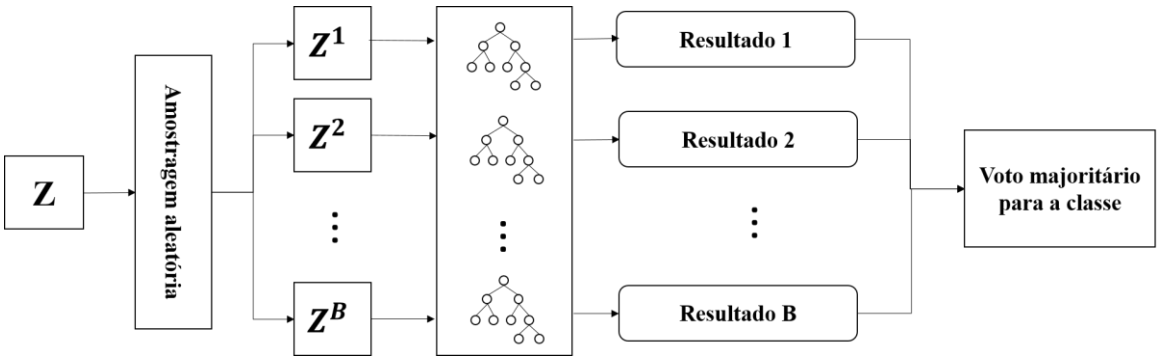
Fonte: HASTIE e outros (2009)

Uma das principais vantagens do algoritmo *Random Forest* em relação às árvores de decisão é que, como prova Breiman (2001), tal algoritmo sofre menos com o problema de sobre ajustes dos dados. Behnia e Blais-stevens (2018) ressaltam que o algoritmo pode ser utilizado em diversas tarefas como classificação, regressão, estimação de densidade, aprendizado semi-supervisionado ou ainda aprendizagem múltipla. Além disso, não é necessária nenhuma

hipótese a respeito da relação entre as variáveis explanatórias e a variável resposta (KIM *et al*., 2018)

Basicamente, o algoritmo pode ser explicado em 3 etapas. A primeira, representa a amostragem, na qual B amostras serão criadas aleatoriamente através de amostragem com reposição (outra denominação que pode ser encontrada é *bootstrapping*), B representa o número de árvore na floresta aleatória. Após isso, na segunda etapa, para cada amostra Z* será criada uma árvore de decisão, seguindo as regras apresentadas no Quadro 2. Após cada uma das B árvores serem criadas, o algoritmo *Random Forest* consistirá na coleção de cada uma dessas árvores de decisão. Por fim, para realizar as predições, é retornada a média das previsões para a problemática de regressão ou a classe cujo se obteve a maior quantidade de votos das árvores do modelo, para a problemática de classificação. No Fluxograma 2 está ilustrada o funcionamento do algoritmo *Random Forest* para a problemática de classificação.

Fluxograma 3 – Princípio de classificação do Random Forest



Fonte: adaptado de Wang e outros (2015c)

Wang e outros (2015c) descreveram o procedimento para geração das árvores de decisão em quatro etapas, como segue:

- **Etapa 1**: Os dados de treino de cada árvore será a amostra gerada aleatoriamente com N observações. O tamanho da amostra de treino de cada árvore é o mesmo que do banco de dados completo, porém, gerado a partir de amostragem com reposição;
- **Etapa 2**: Dado que existem M variáveis condicionantes, uma quantidade m << M será escolhida aleatoriamente para serem testada em cada nó. (BREIMAN, 2001) sugere que o valor inicial seja $m = \sqrt{M}$, porém esse valor pode ser modificado conforme o problema em questão;

- **Etapa 3**: Cada variável deve ser testada e o melhor corte deve ser selecionado através do indicador de impureza, descrito na Equação (3.7). Onde $p(j|t)$ é a probabilidade da classe j no nó t.

$$Gini(x) = 1 - \sum_{j=1}^{k} [p(j|t)]^2 \quad (3.7)$$

- **Etapa 4:** Cada árvore cresce até o tamanho máximo, sem punição.

Para Zhang e outros (2017a) os parâmetros chaves são a quantidade de árvores na floresta (B) e a quantidade de variáveis escolhidas aleatoriamente em cada nó (m). Behnia e blais-stevens (2018) explica que o aumento do número de árvore na floresta resulta na diminuição do erro *out-of-bag* (OOB) do modelo até um determinado limite, a partir do qual não é interessante adicionar mais árvores. Já Gislason e outros (2006) afirmam que ao diminuir a quantidade de variáveis usadas a cada corte, a complexidade computacional do modelo reduz, e a correlação entre as árvore da floresta também diminui.

Como mencionado anteriormente, um conceito importante são os dados OOB. No início do processo de criação de cada árvore, é aplicado um processo de amostragem com reposição, onde aproximadamente 2/3 dos dados originais são usados para gerar o classificador (dados *in-bag*), os demais, 1/3, são chamados de dados OOB, que são usados para validação do modelo Harris e outros (2015). Para estimar a acurácia do modelo, a variável resposta das amostras OOB são preditas com a árvore na qual aquela amostra é OOB e a acurácia calculada através de validação cruzada, comparando o valor predito com o verdadeiro valor (GISLASON *et al*., 2006).

Uma outra característica útil é o cálculo da importância das variáveis. Wang e outros (2015c) explica que no geral há dois métodos para calcular tal índice. O primeiro calcula o erro OOB para cada árvore (OOB_1), então adiciona ruídos aos dados da variável i e recalcula o erro OBB (OOB_2). A importância da variável i é calculada tomando a média da diferença entre os erros OOB_1 e OOB_2 e normalizando com o desvio padrão. O segundo método usa o decréscimo da impureza de cada nó. Sempre que um corte for feito na variável i, a impureza dos nós filhos será menor que a do nó pai, dessa forma, combinado o decréscimo para a variável i em todas as árvores da floresta fornece uma medida da importância da variável.

Calle e Urrea (2011) discutem dois métodos implementados no pacote *randomForest*, na linguagem R, que usam o decréscimo da impureza para calcular a importância das variáveis. O primeiro é o decréscimo médio da acurácia (*mean decrease accuracy* - MDA). O MDA mesura a importância de cada variável através do cálculo da variação na acurácia da predição quando os valores das variáveis são permutados aleatoriamente, comparados com os valores originais.

O segundo utiliza o decréscimo médio do índice GINI (*mean decrese gini* – MDG). Já o MDG calcula a importância da variável i através da soma de todos os decréscimos no índice GINI devido ao corte na variável i, normalizada pelo número de árvores na floresta. Ambas as medidas (MDA e MDG) serão utilizadas no presente trabalho.

3.4.3 Avaliação de performance

Avaliação da performance é uma das mais importantes etapas no desenvolvimento de qualquer solução baseada em Aprendizado de Máquina (RAMASUBRAMANIAN; SINGH, 2017). Desde que tenha um modelo final, entende-se por avaliação da performance a tarefa de estimar o erro da predição em novos dados, ou seja, estimar o erro da generalização (HASTIE *et al.*, 2009).

A ideia central da avaliação do modelo é minimizar o erro nos dados de teste, onde o erro pode ser definido de várias maneiras. De maneira mais intuitiva, erro pode ser definido como a diferença entre o valor atual de uma variável resposta e o valor predito pelo modelo de aprendizado de máquina (RAMASUBRAMANIAN; SINGH, 2017). Ainda segundo Ramasubramanian e Singh (2017), a métrica de avaliação do erro usada para treinar o modelo deve ser diferente da métrica usada para avaliar o modelo.

Avaliação da performance do modelo pode ser realizada uma vez que tenha desenvolvido um modelo e tem-se o desejo de entender qual o desempenho nos dados de teste e validação. Antes de mais nada, é comum dividir o banco de dados em três categorias (RAMASUBRAMANIAN; SINGH, 2017):

- **Dados de treino**: esses dados serão usados para treinar o modelo, de modo a otimizar alguma métrica pré-definida para o ajuste do modelo;
- **Dados de teste**: Tais dados contém pontos que o algoritmo de Aprendizado de Máquina não usou na etapa de treino. Esse banco de dados é utilizado para verificar como o modelo desempenha a tarefa em novos dados, e verificar se é necessária alguma melhoria;
- **Dados de validação**: em muitos casos, esse banco de dados não é criado, por diversas razões como: limitação dos dados; pouco tempo disponível; banco de dados de teste muito grande. O principal objetivo desse banco de dados é identificar sobre ajustes (*overfitting*) nos dados e prover ideia para necessidades de calibração.

Segundo Hastie e outros (2009) é difícil fornecer uma regra geral para a escolha do proporção ideal de cada uma das três partes. Tipicamente é usada a proporção 2:1:1.

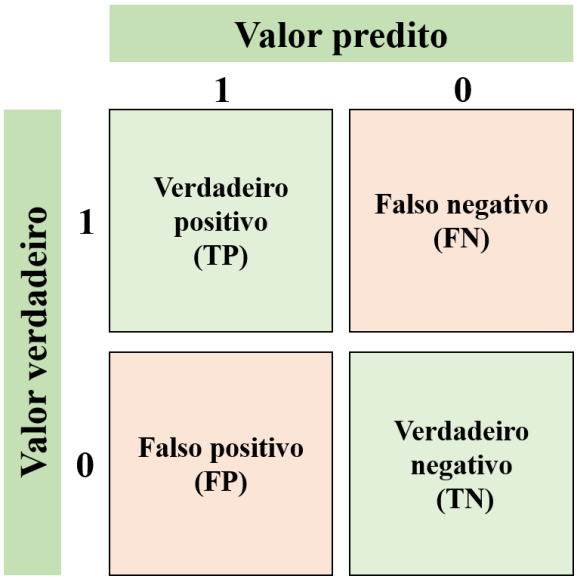
O método utilizado para avaliar a performance irá depender, entre outras coisas, do tipo de saída do modelo, se é uma saída contínua, como regressão, por exemplo, ou saída discreta, como problemas de classificação (RAMASUBRAMANIAN; SINGH, 2017). No presente trabalho todas as saídas são discretas, por isso, a partir de agora todos os conceitos serão relacionados a tal tipo de saída.

Para Ramasubramanian e Singh (2017) existem 3 objetivos principais da avaliação de performance de um modelo. A acurácia do modelo que reflete a proporção de predições corretas. O ganho do modelo, ou seja, compara a saída do modelo com os resultados que seriam obtidos sem usar o modelo ou usar um modelo aleatório. Por fim, a credibilidade do modelo, que reflete a segurança em usar o modelo em dados diferentes dos utilizados durante a fase de treinamento.

3.4.3.1 Validação cruzada

Um dos conceitos mais usados na avaliação e performance é a validação cruzada com o uso da matriz confusão e suas métricas derivadas. A Figura 4 exibe a organização de uma matriz confusão.

Figura 4 – Matriz confusão



Fonte: O Autor (2019)

Lantz (2015) explica que a classe de interesse é conhecida como classe positiva. Já a classe da qual não temos interesse é conhecida como classe negativa. Seguindo tal definição e

quando a matriz confusão tem dimensão 2x2 é possível classificar a predição em quatro categorias:

- Verdadeiro Positivo (TP): Corretamente classificado na classe de interesse;
- Verdadeiro negativo (TN): corretamente classificado na classe sem interesse;
- Falso Positivo (FP): Incorretamente classificado na classe de interesse;
- Falso Negativo (FN): Incorretamente classificado na classe sem interesse.

Apesar da Figura 4 mostrar apenas duas classes, uma matriz confusão pode ter quantas classes forem necessárias. Nesse caso, os elementos da diagonal principal representam os acertos (TP + TN), os demais valores representam os erros da predição (LANTZ, 2015).

Com base na matriz confusão, é possível definir as medidas de desempenho expostas na Equação (3.8) a Equação (3.10).

$$acurácia = \frac{TP + TN}{Total\ de\ observações} \tag{3.8}$$

$$sensibilidade = \frac{TP}{TP+FN} \tag{3.9}$$

$$especificidade = \frac{TN}{TN + FP} \tag{3.10}$$

Ramasubramanian e Singh (2017) fornece uma definição clara para as medidas sensibilidade e especificidade. Entende-se por sensibilidade a probabilidade de o teste indicar verdadeiro positivo entre os valores verdadeiramente positivos, também conhecida como taxa de verdadeiro positivos. Já especificidade é entendida como a probabilidade de o teste indicar os casos verdadeiros negativos entre os casos verdadeiramente negativos, também conhecido como taxa de verdadeiro negativos ou *recall*.

Fielding e Bell (1997) fornece ainda outras medidas que podem ser derivadas da matriz confusão, expostas no Quadro 3.

O índice Kappa (K) é amplamente utilizado para o estudo de confiabilidade de sistemas de classificação de dados categóricos (PERROCA; GAIDZINSKI, 2003). Pode ser definido como uma medida de concordância entre classificações realizadas, comparando o valor real com o predito. (LANDIS; KOCH, 1977) estabeleceram intervalos com os quais é possível avaliar a qualidade da classificação de acordo com K, conforme exposição da Tabela 1.

Quadro 3– Medidas de performance derivadas da matriz confusão

Medida	Fórmula
Prevalencia	(TP + FN)/ N
Poder geral de diagnóstico	(FP + TN)/N
Taxa de falsos positivos	FP/(FP + TN)
Taxa de falsos negativos	FN/(TP+FN)
Poder de predição de positivos	TP/(TP+FP)
Poder de predição de negativos	TN/(FN+TN)
Taxa de classificação incorreta	(FP + FN)/N
Índice Kappa	[(TP + TN) – (((TP + FN)(TP + FP) + (FP + TN)(FN + TN))/N)]/[N – (((TP + FN)(TP + FP) + (FP +TN)(FN+TN))/N)]

Fonte: (FIELDING; BELL, 1997)

Tabela 1 – Intervalos do índice Kappa

Statística Kappa	Concordância
K < 0	Pobre
0 > K ≤ 0,20	Fraca
0,20 > K ≤ 0,40	Razoável
0,40 > K ≤ 0,60	Moderada
0,60 > K ≤ 0,80	Alta
0,80 > K ≤ 1	Quase perfeita

Fonte: (LANDIS; KOCH, 1977)

É comum em problemas de aprendizado supervisionado, lidar com banco de dados desbalanceados. Um banco de dados está desbalanceado quando a quantidade de casos da classe negativa presentes no banco de dados de treino é muito maior que da classe positiva (CASTRO; BRAGA, 2012). Sendo assim, é necessário modificar as métricas de avaliação de tais problemas, pois a medida de acurácia não consegue detectar os erros nas classes desfavorecidas.

Considere um exemplo de um banco de dados em que a classe minoritária é representada por apenas 2% das observações. Um classificador com acurácia de 98% pode ser obtido simplesmente por classificar todas as observações como pertencentes a classe majoritária. O problema é apesar da taxa de acurácia elevada, o modelo não foi capaz de classificar nenhuma observação de interesse corretamente.

Para avaliar modelos com tais características, Castro e Braga (2012) sugerem utilizar métricas de medição de performance que façam distinção entre os erros cometidos para cada

classe. Dentre as métricas sugeridas, os autores recomendam a utilização das medidas derivadas da matriz confusão: taxas de falso positivo; taxa de falso negativo; taxa de verdadeiros positivos e; taxa de verdadeiros negativos.

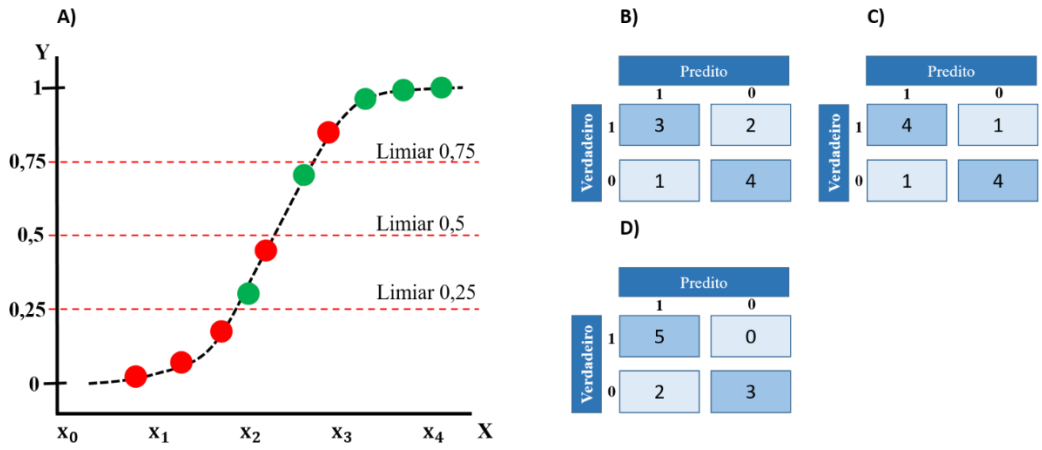
3.4.3.2 Área sobre a curva ROC

A curva de características operacionais do receptor (*Receiver Operating Characteristic* – ROC) é uma representação gráfica da performance de classificadores binários com mudanças no limiar de corte para classificar uma determinada observação (RAMASUBRAMANIAN; SINGH, 2017). O nome da técnica (*Receiver Operating Characteristic*) se refere a performance ("*Operating Characteristic*") de um observador humano, máquina ou algoritmo ("*Receiver*") na tarefa de classificar um evento em classes dicotômicas, como por exemplo: normal ou anormal, negativo ou positivo; saudável ou doente (DELEO, 1993).

A curva ROC é criada através da relação entre a taxa de verdadeiros positivos (sensibilidade) no eixo y e a taxa de falsos positivos (1 – especificidade) equivalente no eixo x. Cada ponto representa um limiar de classificação diferente (FIELDING; BELL, 1997).

Para exemplificar o procedimento descrito por Fielding e Bell (1997), veja exemplo fictício exposto na Figura 5. Tal problema classifica as observações em duas classes dicotômicas: positivo e negativo. Os pontos verdes representam os pontos verdadeiramente positivos, enquanto os pontos vermelhos representam os pontos verdadeiramente negativos. Para classificar as observações foi utilizado um modelo de regressão logística, o qual classifica as observações na classe positivo caso a probabilidade seja maior que o limiar escolhido, representado pelas linhas vermelhas na Figura 5 (a).

Figura 5– Problema fictício de classificação logística com A) representação gráfica e; matrizes confusão para os limiares de B) 0,75; C) 0,5; D) 0,25

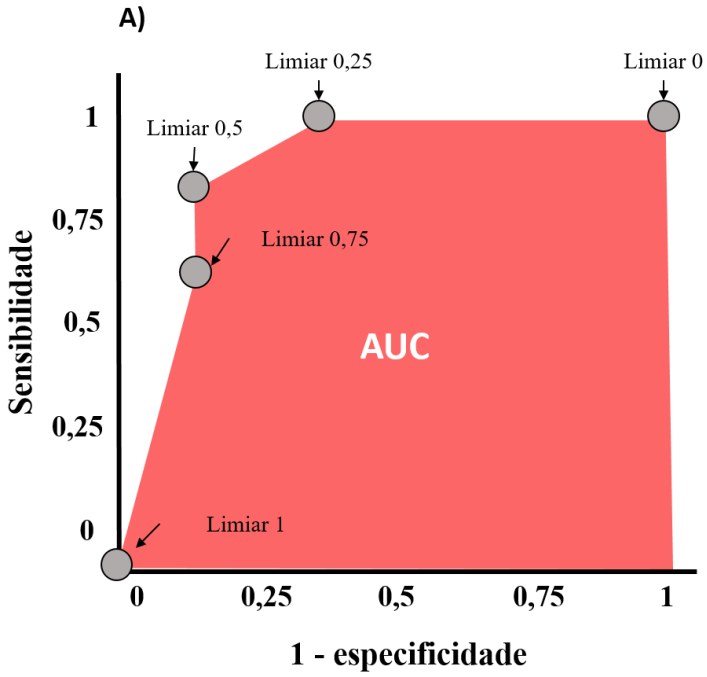


Fonte: O Autor (2019)

Perceba que quando o limiar de classificação é igual a 0,75 as observações serão classificadas na classe positivo caso a probabilidade seja maior ou igual a 0,75. Isso faz com que 3 dos 5 pontos verdadeiramente positivos sejam classificados como positivos, enquanto 2 deles foram classificados como negativos. Da mesma forma, dos 5 pontos verdadeiramente negativos, 4 deles foram classificados como negativos e apenas 1 foi classificado como positivo. Com esses dados é possível construir a matriz confusão exposta na Figura 5 (b). Da mesma forma descrita anteriormente, ao mudar o limiar é possível construir uma matriz confusão para cada limiar, a Figura 5 (c) representa a matriz confusão quando o limiar é igual a 0,5. A Figura 5 (d) exibe a matriz confusão quando o limiar é igual a 0,25.

A curva ROC resume as informações descrita no parágrafo anterior. Para isso, como mostra a Figura 6, para cada matriz confusão é calculada a sensibilidade utilizando a Equação (3.8) e a especificidade, utilizando a Equação (3.9). Cada ponto representa um limiar da classificação e as respectiva sensibilidade e taxa de falsos positivos (1 - especificidade).

Figura 6– Curva ROC



Fonte: O Autor (2019)

Deleo (1993) afirma que a Área Sobre a Curva ROC (AUC) é uma medida importante para a análise dos modelos, visto que fornece uma medida simples para a acurácia geral. O valor da AUC varia entre 0,5 e 1. Se o valor for igual a 0,5, isso indica que não há diferença entre as

duas classes, enquanto se o valor for igual a 1 indica que não há sobreposição da distribuição das classes (FIELDING; BELL, 1997).

Valores iguais a 1 dificilmente serão alcançados (FIELDING; BELL, 1997). AUC igual a 0,8 significa que para 80% das vezes em que uma seleção aleatória for realizada na classe positiva terá mais sucessos que uma seleção aleatória conduzida na classe negativa (DELEO, 1993).

A Curva ROC, além de fornecer informações sobre a acurácia do modelo, serve como ferramenta para desenvolver regras de decisão (DELEO, 1993). Segundo Fielding e Bell (1997), são necessários dois elementos para identificar os limiares mais adequados. Um dos pontos necessários é a identificação dos custos referentes a falsos positivos e de falsos negativos. Tais custos podem ser difíceis de serem levantados e variam bastante a depender do problema em questão (FIELDING; BELL, 1997). Como guia, Fielding e Bell (1997), sugerem que se os custos de falsos positivos sejam maiores que os custos de falsos negativos, o limiar deve favorecer a especificidade, caso contrário a sensibilidade de ser favorecida.

A versatilidade de tal técnica é um dos fatores que levaram a sua ampla utilização. Shrestha e outros (2017) utilizaram tal método para avaliar a performance do algoritmo *Random Forest* na tarefa de avaliação da susceptibilidade a deslizamentos de terras. Da mesma forma Althuwaynee e outros (2014) avaliaram a performance de modelos de regressão logística. Tehrany e outros (2015a) avaliaram a performance do algoritmo SVM na tarefa de avaliação do perigo de inundações.

3.5 Conclusões do capítulo

No presente capítulo foram apresentados e discutidos os conceitos utilizados no presente estudo. Em primeiro lugar, os termos referentes ao gerenciamento de risco foram apresentados e discutidos, diferenciando cada um deles. Além dos termos citados anteriormente, os conceitos referentes a desastres naturais foram definidos e caracterizados. Os algoritmos e métodos utilizados foram apresentados, bem como as métricas de validação de performance. Tal capítulo foi responsável por formar a base conceitual necessária para o completo entendimento do presente trabalho.

4 REVISÃO SISTEMÁTICA DA LITERATURA

O presente capítulo tem como propósito apresentar a revisão sistemática da literatura, que tem como objetivo prospectar trabalhos científicos publicados em periódicos de alto impacto no tema de mapeamento de perigo de deslizamentos e inundações utilizando algoritmos de aprendizado de máquina, tais trabalhos serão utilizados para responder questões chaves que permitirão identificar padrões e traçar novas perspectivas para o tema. Para a seleção desses trabalhos foi aplicada uma metodologia sistemática de busca e filtragem. Após a busca, todos os trabalhos foram analisados com o objetivo de responder as questões de pesquisa previamente definidas.

4.1 Metodologia

Essa seção tem como propósito explicar a metodologia utilizada no capítulo para analisar a literatura científica sobre mapeamento do perigo de inundações e deslizamentos com a utilização de algoritmos de aprendizado de máquina. Para Kitchenham e outros (2007), uma revisão sistemática da literatura é um meio para identificar, avaliar e interpretar todas as pesquisas relevantes para uma questão de pesquisa específica, um campo de estudo, ou um fenômeno de interesse. Gupta e outros (2018) afirmam que para uma revisão ser sistemática, deve-se responder a questões específicas e aplicar uma metodologia clara para a avaliação de todas as informações disponíveis.

Ainda segundo Kitchenham e outros (2007), existem várias razões para se realizar uma revisão sistemática da literatura: (i) Para resumir as evidências existentes referentes a um método ou tecnologia; (ii) Para identificar oportunidades de pesquisas futuras; (iii) ou ainda para propor estruturas ou base de conhecimento para sustentar novas atividades de pesquisa.

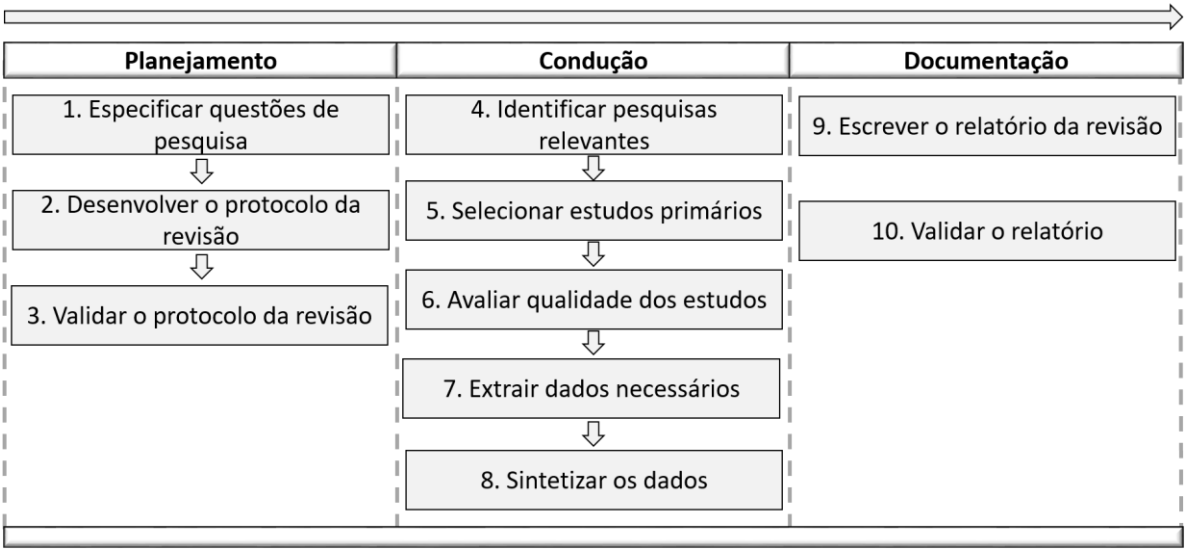
Brereton e outros (2007) propõem que o processo de revisão sistemática da literatura seja dividido em dez etapas, segmentadas em 3 fases, como mostra o Fluxograma 4.

Já para (GUPTA *et al*., 2018), o processo de revisão sistemática da literatura se divide em 14 etapas, que são:

1. Desenvolver as questões de pesquisa;
2. Avaliar a qualidade das questões de pesquisa;
3. Estabelecer os critérios de inclusão;
4. Desenvolver o protocolo do estudo;
5. Registrar a revisão;
6. Selecionar bancos de dados;

7. Conduzir a busca;
8. Avaliar a qualidade da pesquisa;
9. Filtrar estudo;
10. Extrair dados;
11. Avaliar viés no estudo;
12. Analisar os dados;
13. Sintetizar e interpretar resultados;
14. Reportar resultados.

Fluxograma 4– Processo de revisão sistemática da literatura

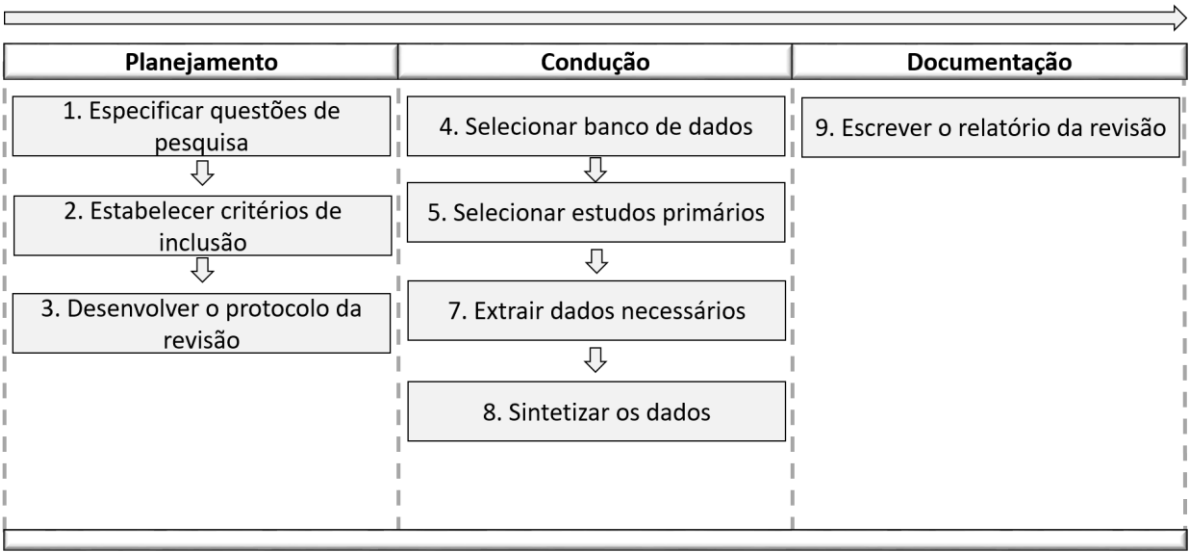


Fonte: adaptado de (BRERETON et al., 2007)

Os trabalhos publicados por Brereton e outros (2007), Kitchenham e outros (2007) e Gupta e outros (2018) possuem certo grau de consenso sobre as etapas necessárias e características essenciais para a realização de uma boa revisão sistemática da literatura. Assim sendo, a metodologia utilizada no presente trabalho tomou como base os trabalhos citados anteriormente, como mostra o Fluxograma 5.

Apesar da forma sequencial na qual a metodologia adotada foi apresentada, vale ressaltar que o processo é interativo e podem ocorrer interações fora da ordem pré-estabelecida. Nas próximas subseções, cada uma das fases apresentadas no Fluxograma 5 serão detalhadas.

Fluxograma 5– Metodologia adotada no estudo



Fonte: O Autor (2019)

4.1.1 Questões de pesquisa

Kitchenham e outros (2007) sugerem que as questões de pesquisa devam ser elaboradas informando a população a ser estudada, o meio de intervenção, a forma de comparação e as saídas a serem analisadas. Para o presente trabalho, a população a ser estudada são os métodos de aprendizado de máquina. Já a intervenção é o campo de estudo referente a mapeamento do perigo de deslizamentos e inundações. A forma de comparação será entre os artigos selecionados para o estudo. Por fim, a saída específica, será definida em cada questão de pesquisa.

Assim sendo, o presente trabalho tem como objetivo responder as seguintes questões de pesquisa:

- **RQ1:** Quais métodos de aprendizado de máquina são mais utilizados para o mapeamento do perigo de enchentes e inundações?
- **RQ2:** Como estão distribuídos os valores da avaliação de performance dos métodos de aprendizado de máquina para o mapeamento do perigo de enchentes e inundações?
- **RQ3:** Quais as variáveis condicionantes mais utilizadas nos métodos de aprendizado de máquina para o mapeamento do perigo de enchentes e inundações?

- **RQ4:** Como estão distribuídos os tamanhos das amostras utilizadas nos métodos de aprendizado de máquina para o mapeamento do perigo de enchentes e inundações?

4.1.2 Estabelecer critérios de inclusão

Segundo Gupta e outros (2018), os critérios de inclusão devem ser claros o suficiente para tornar possível a identificação do tipo de estudo a ser analisado, o que pode incluir: a população estudada, o tipo do estudo considerado, entre outros fatores que caracterizam a área de estudo em específico.

Os critérios de inclusão adotados estão divididos em três grupos, conforme exposição do Quadro 4. O primeiro reflete as características do documento a ser analisado. O segundo, referente a área de estudo. Por fim, o último grupo referente ao tipo do estudo.

Quadro 4– Critério de inclusão

Grupo	Critérios
Características do documento	• Somente Artigos publicados em periódicos; • Somente artigos escritos em língua inglesa; • Somente artigos publicados até 31/12/2018.
Área do estudo	Serão consideradas somente as seguintes áreas de estudo: • Geologia; • Recurso hídricos; • Engenharia; • Ciências ambientais e ecologia; • Ciências meteorológicas; • Geografia física; • Sensoriamento remoto; • Ciência da computação; • Matemática; • Tecnologia outros tópicos; • Pesquisa operacional e ciência da gestão.
Tipo do estudo	• Artigos que mapeiam o perigo de enchentes e deslizamentos de terra utilizando métodos de aprendizado de máquina; • Somente artigos que realizam o mapeamento espacial do perigo.

Fonte: O Autor (2019)

4.1.3 Desenvolver protocolo da revisão

Um protocolo de revisão é o documento responsável por especificar qual método será usado para realizar uma revisão sistemática (KITCHENHAM *et al*., 2007). Para Gupta e outros (2018), um protocolo de revisão bem elaborado facilita o gerenciamento da revisão, além de evitar redundâncias.

Para Kitchenham e outros (2007), um protocolo de revisão deve conter, entre outros:

- Justificativa para a pesquisa;
- Questões de pesquisa;
- Estratégia de busca;
- Critério de inclusão;
- Procedimentos para seleção dos estudos;
- Procedimento de verificação da qualidade dos trabalhos selecionados;
- Estratégia da extração dos dados;
- Síntese dos dados extraídos;
- Estratégia de disseminação.

Cada um desses tópicos, quando pertinente, será explicado nas subseções posteriores.

O protocolo de revisão foi elaborado pelo pesquisador e avaliado pelo orientador dessa pesquisa. O protocolo em sua íntegra encontra-se no Apêndice A.

4.1.4 Selecionar banco de dados e termos de busca

A base de dados selecionada para o presente estudo foi a Web of Science core Collection. Tal base possui mais de vinte mil periódicos disponíveis, todas as referências são indexadas, além de possuir várias informações sobre cada trabalho publicado. Para realizar a busca, primeiro foram definidos os termos de busca e por fim, a busca inicial foi realizada no mecanismo disponibilizado.

Foram definidos dois grupos de termos de busca baseados na literatura específica. O primeiro, relacionado as principais técnicas de aprendizado de máquina utilizadas na literatura. O segundo, referente a termos geralmente utilizados para representar o mapeamento de perigo de enchentes e deslizamentos. Por exemplo, Feng e outros (2016), usaram o termo "*landslide susceptibility*", Muñoz e outros (2018) usaram o termo "*Flood Forecasting*", Kourgialas e Karatzas (2017) usaram "*flood hazard mapping*", e assim por diante. O Quadro 5 resume todos os termos de busca de ambos os grupos.

Todas as buscas foram realizadas nos campos título, resumo, palavras chaves e palavras chaves mais citadas. Os termos do primeiro grupo foram combinados com os termos do segundo grupo com o operador lógico AND, formando, por exemplo, o termo de busca: "*support vector regression*" AND "*Landslid* risk*".

Por fim, as buscas foram realizadas, resultando em 1471 trabalhos. Após essa etapa, os filtros definidos anteriormente foram aplicados para enfim chegar na amostra utilizada no estudo.

Quadro 5– Termos de busca utilizados

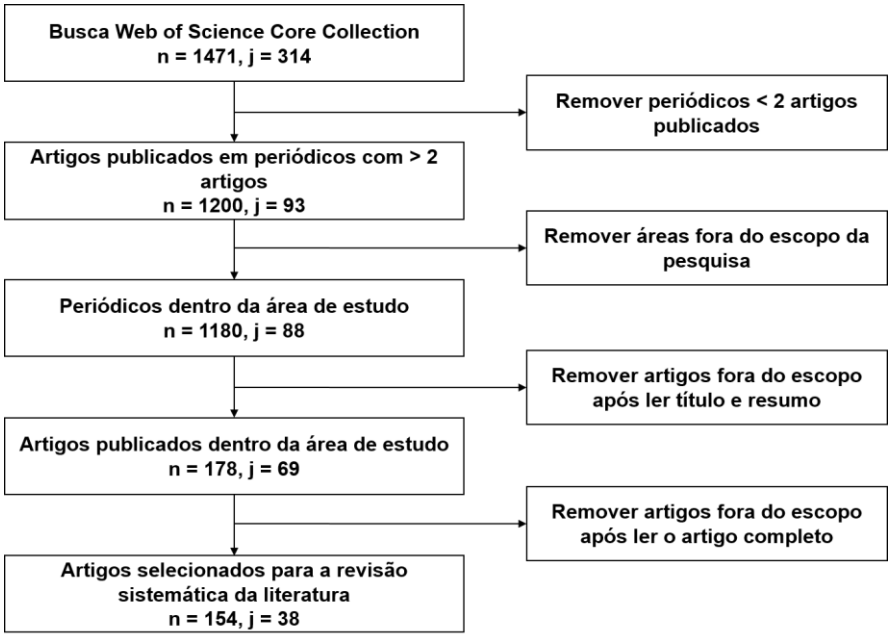
Grupo relacionado a Aprendizado de Máquina	**Grupos relacionados aos desastres**
"Machine Learning" OR "Deep-Learning" OR "data mining" OR "artificial intelligence" OR "nearest neighbo" OR "K-NN" OR "decision tree*" OR " linear regression" OR "regression tree*" OR "classification trees" OR "neural network*" OR "ANN" OR "genetic algorithm" OR "association rule*" OR "support vector machine*" OR "SVM" OR "support vector regression" OR "random forest" OR "boosting"; OR "ensemble learning" OR "ensemble model*" OR "gradient descent" OR "clustering" OR "logistic regression" OR "genetic algorithm" OR "naive bayes" OR "bagging" OR "Least-square support vector machines" OR "K-Means" OR "Dimensionality Reduction" OR "boosting" OR "adaboost" OR "Principal component analysis" OR "Classifcation and Regression Tree" OR "classifcation rules" OR "Association Rules" OR "Linear Discriminant Analysis"*	"*Flood* prediction" OR "Flood* vulnerability" OR "Flood* estimation" OR "Flood* forecast" OR "Flood* analysis" OR "Flood* susceptibility" OR "Flood* assessment"OR "Flood* hazard" OR "Flood* risk" OR* "Landslid * *vulnerability" OR* "Landslid* *prediction" OR* "Landslid* *estimation" OR* "Landslid* *forecast" OR* "Landslid* *analysis" OR "*Landslid* *susceptibility" OR "*Landslid* *assessment"OR* "Landslid* *hazard" OR* "Landslid* *risk" OR* "inundation * *prediction" OR* "inundation *estimation" OR* "inundation *forecast" OR* "inundation *analysis" OR* "inundation * *vulnerability" OR "*inundation *susceptibility" OR "*inundation *assessment" OR* "inundation *hazard" OR* "inundation *risk"*

Fonte: O Autor (2019)

4.1.5 Selecionar estudos primários

Tendo o resultado da busca inicial, alguns filtros e critérios foram aplicados para definir a amostra que será utilizada no estudo. A busca inicial resultou em 1471 trabalhos, distribuídos em 314 periódicos, mostrando a variedade de estudos no tema em questão. Após aplicar todos os critérios estabelecidos, foi possível definir a amostra que será utilizada no estudo, contendo 154 artigos, distribuídos em 38 periódicos. Todo o processo de seleção dos estudos primários está exposto no Fluxograma 6.

Fluxograma 6– Processo de seleção dos estudos primários



Fonte: O Autor (2019)

4.1.6 Extrair os dados necessários

Os dados foram extraídos com o auxílio do pacote desenvolvido na linguagem R: Bibliometrix Aria e Cuccurullo (2017). Tal ferramenta fornece um banco de dados com 37 variáveis diferentes, contendo informações sobre cada trabalho incluído na pesquisa, dentre elas:

- AU: Autores;
- TI: Título;
- SO: Fonte;
- JI: Abreviação ISSO para a fonte;
- DT: Tipo do documento;
- DE: Palavras chaves;
- AB: Resumo;
- C1: Nacionalidade do autor;
- CR: Referências citadas;
- TC: Número de citações;
- PY: Ano;
- SC: Categoria.

Além das análises proporcionadas por tal ferramenta, serão extraídas as seguintes informações complementares, totalizando 42 informações coletadas:

- Método de aprendizado de máquina utilizado com maior performance;
- Valor numérico da performance;
- Variáveis condicionantes adotadas no modelo;
- Tamanho da amostra utilizada;
- Maneira de validação do modelo.

O método de aprendizado de máquina com maior performance diz respeito ao algoritmo utilizado no estudo que alcançou maior valor numérico na métrica de performance utilizada pelos autores. Já performance corresponde ao valor numérico do algoritmo eu melhor desempenhou a função descrita no estudo. As variáveis condicionantes dizem respeito a quais variáveis independentes os autores escolheram para utilizar no estudo. O tamanho da amostra informa sobre a quantidade de eventos identificados previamente e utilizados para o treinamento dos algoritmos. Por fim, a validação do modelo diz respeito aos métodos utilizados pelos autores para compararem os resultados gerados pelo modelo com a dinâmica real dos eventos.

4.1.7 Síntese e escrita do relatório

A síntese e escrita serão apresentadas nas próximas seções. Na seção 4.2 serão demonstrados os resultados das análises realizadas. Em seguida, será proposto um processo para o mapeamento de perigo de inundações e deslizamentos usando técnicas de aprendizado de máquina.

A síntese dos dados, como mencionado anteriormente será realizada com o auxílio de duas ferramentas. A primeira é o pacote, escrito na linguagem R, Bibliometrix (ARIA; CUCCURULLO, 2017). Tal pacote fornece uma vasta quantidade informações sobre os estudos preliminares de forma rápida e confiável. Além dessa ferramenta, será utilizado também planilhas eletrônicas para coletar informações mais específicas.

O banco de dados formado contém 42 colunas e 178 linhas. Os dados brutos foram armazenados para fins de verificação. Por fim, foram utilizados pacotes de manipulação e visualização de dados disponíveis na linguagem R.

4.2 Resultados e discussões

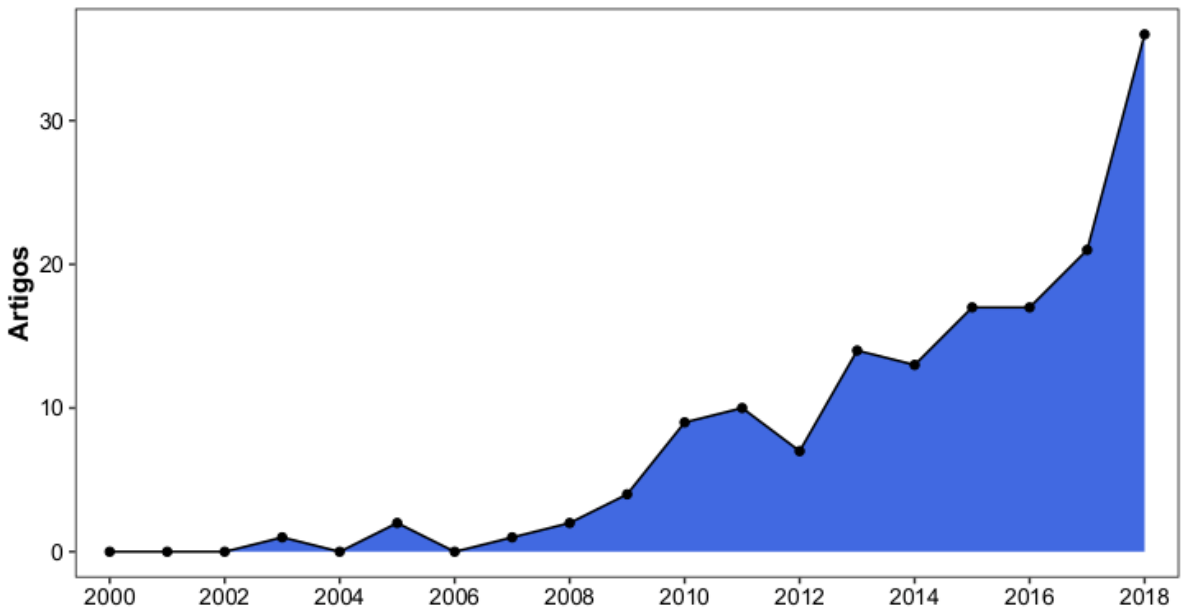
A presente seção irá apresentar a análise exploratória dos dados. Em primeiro lugar, uma análise geral dos dados será realizada, mostrando informações como: evolução das pesquisas;

principais autores; principais fontes; entre outras. Por fim, cada uma das questões de pesquisa será sistematicamente respondida.

4.2.1 Análise exploratória da amostra

Como indício da relevância e atualidade do tema, o número de publicações apresentou tendência crescente, como mostra o Gráfico 1. Como evidência, a taxa de crescimento anual apresentou valor de 31,74% ao ano. Nota-se um crescimento acentuado nos números de publicações a partir de 2010, fato ocorrido provavelmente pela larga popularização das técnicas de aprendizado de máquina, bem como o surgimento de tecnologia que possibilitaram a análise de grandes volumes de dados.

Gráfico 1– Evolução histórica dos trabalhos



Fonte: O Autor (2019)

O relatório publicado por WIPO (2019), apresenta relevantes contribuições a respeito das tendências em tecnologias que utilizam a inteligência artificial. Segundo tal relatório, as áreas com maior crescimento foram transporte, agricultura e aplicações governamentais, tais como gerenciamento de desastres, as quais possuem uma taxa de crescimento de no mínimo 30% ao ano, tendo a China como o país líder em utilização de tais tecnologias.

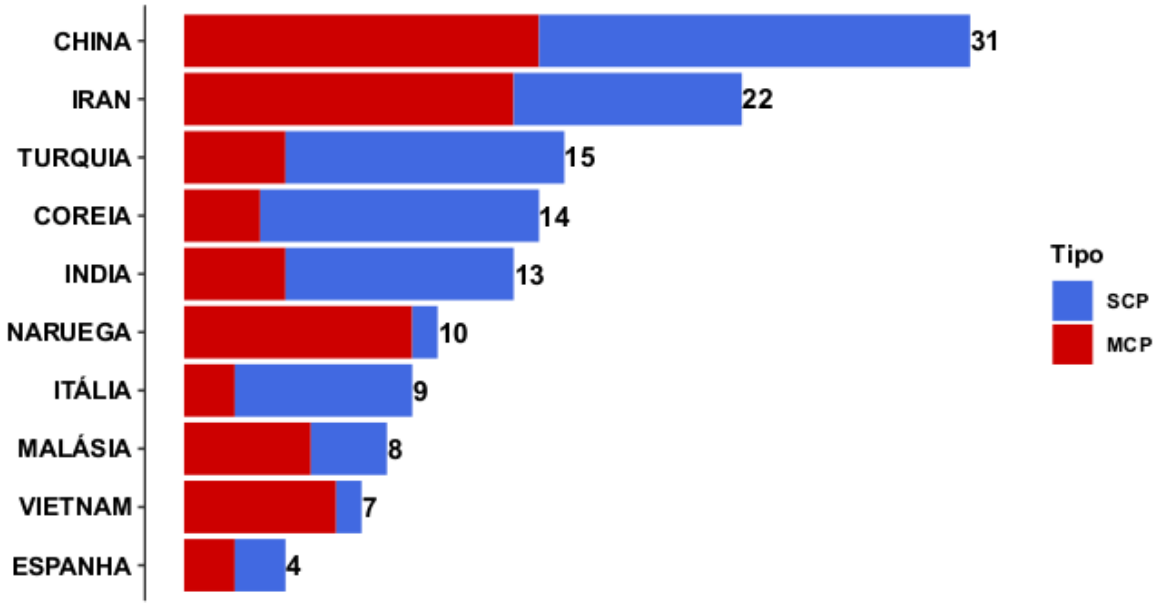
Ainda segundo WIPO (2019), a área de aprendizado de máquina é uma das principais responsáveis pelo grande crescimento das tecnologias de inteligência artificial. As áreas com

crescimento constante nos últimos anos foi o NLP e análise preditiva. Segundo WIPO (2019), o crescimento acentuado nas aplicações de inteligência artificial começou nos últimos 7 anos, movido principalmente pelo crescimento do poder computacional e a grande conectividade com grandes volumes de dados compilados e compartilhados.

Segundo os dados analisados na amostra, a China foi o país com maior número de publicações, seguida por Iran e Turquia. Além disso, há um grande número de artigos publicados nesses países com colaboração de autores de outros países, como mostra o Gráfico 2.

Dos dez países expostos no Gráfico 2, seis deles estão entre os 20 países mais afetados por inundações e deslizamentos, segundo dados do EM-DAT (2019). Segundo tais dados, a China é o país mais afetado em danos financeiros por enchentes e deslizamentos. Além da China, a Índia, Itália, Coréia e Iran estão entre os 20 países mais afetados por tais desastres naturais.

Gráfico 2– Produção por país

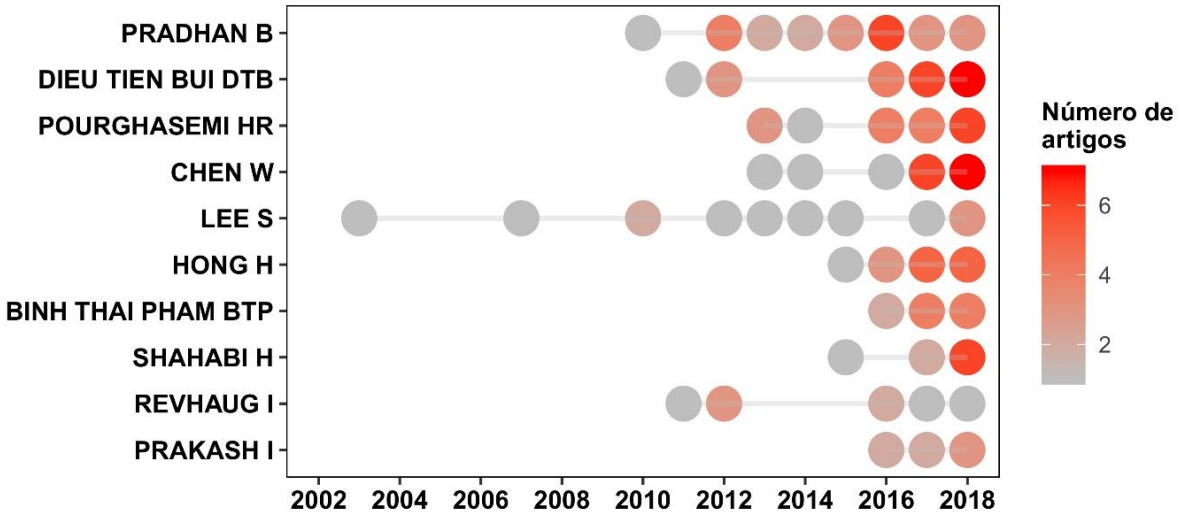


SCP: Publicações com autores de único país; MCP: Publicações com autores de multiplos países

Fonte: O Autor (2019)

Foram identificados 375 autores diferentes, o que significa 2,44 autores por documento. No Gráfico 3 estão expostos os dez autores com maior número de publicação no tema. Destaque para Pradhan B., que entre 2010 e 2018 publicou 24 artigos.

Gráfico 3– Autores mais produtivos



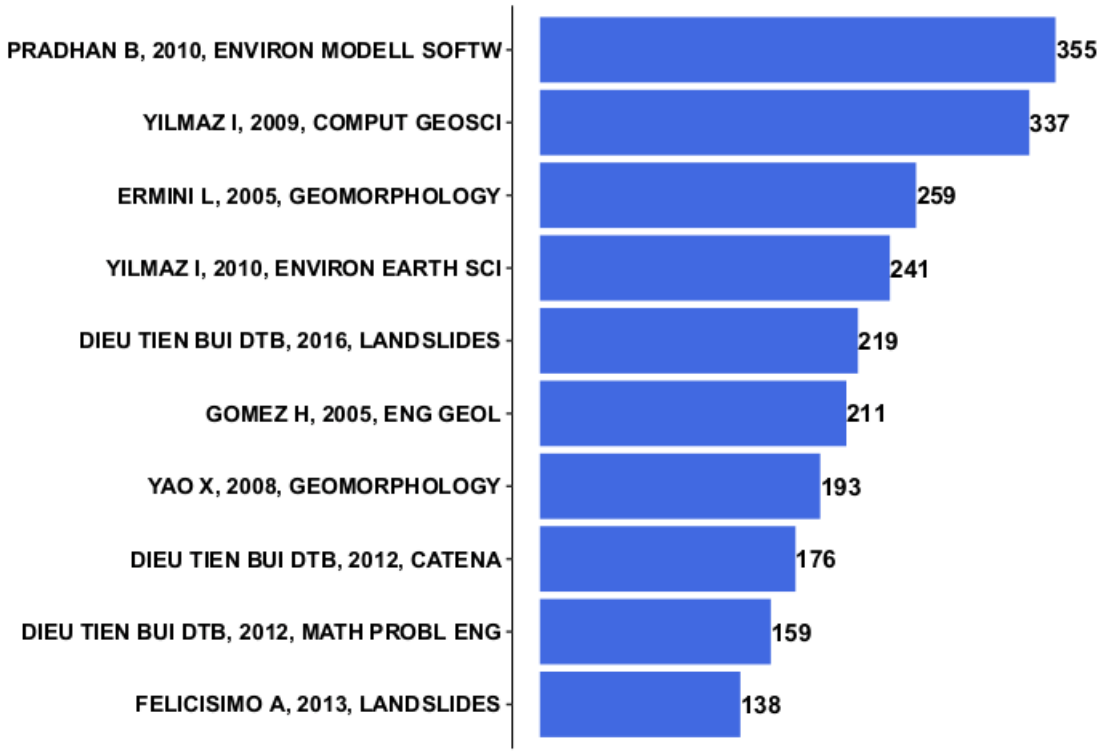
Fonte: O Autor (2019)

Como exposto no Gráfico 4, o trabalho mais citado foi o publicado por Pradhan e Lee (2010), no qual os autores comparam os modelos de Redes Neurais Artificias (ANN), *Frequency Ration* e Regressão logística na tarefa de mapeamento do perigo de deslizamento. Vale ressaltar que os números de citações não devem ser tomados como referência absoluta para classificar a produtividade do autor, pois é necessário levar em consideração o tempo de publicação de cada artigo.

Segundo a lei de Lotka ou Lei do Quadrado Inverso, um número restrito de pesquisadores produz muito em uma determinada área, enquanto um grande volume de pesquisadores produz pouco Machado Junior e outros (2016). A implicação é que a quantidade de autores que publicam n artigos é igual a $1/n^2$ da quantidade de autores que publicam somente 1 artigo. É chamado de coeficiente de Lotka o expoente de n. Para a distribuição ideal o coeficiente é igual a 2, dessa forma, para comparar uma amostra com a distribuição ideal basta comparar o coeficiente de Lotka.

Ao analisar a produtividade da área segundo a Lei de Lotka Machado Junior e outros (2016), foi concluído que os autores possuem produtividade acima do padrão teórico. A amostra apresentou coeficiente de Lotka igual a 1,72. O teste de hipótese para a igualdade entre a distribuição real e a teórica retornou valor-p de 0,1813. Isso significa que a diferença entre o teórico e o observado não é estatisticamente significativa, mostrando que a produção dos autores é igual ao padrão teórico estabelecido, configurando-os como produtivos.

Gráfico 4– Citações por autor



Fonte: O Autor (2019)

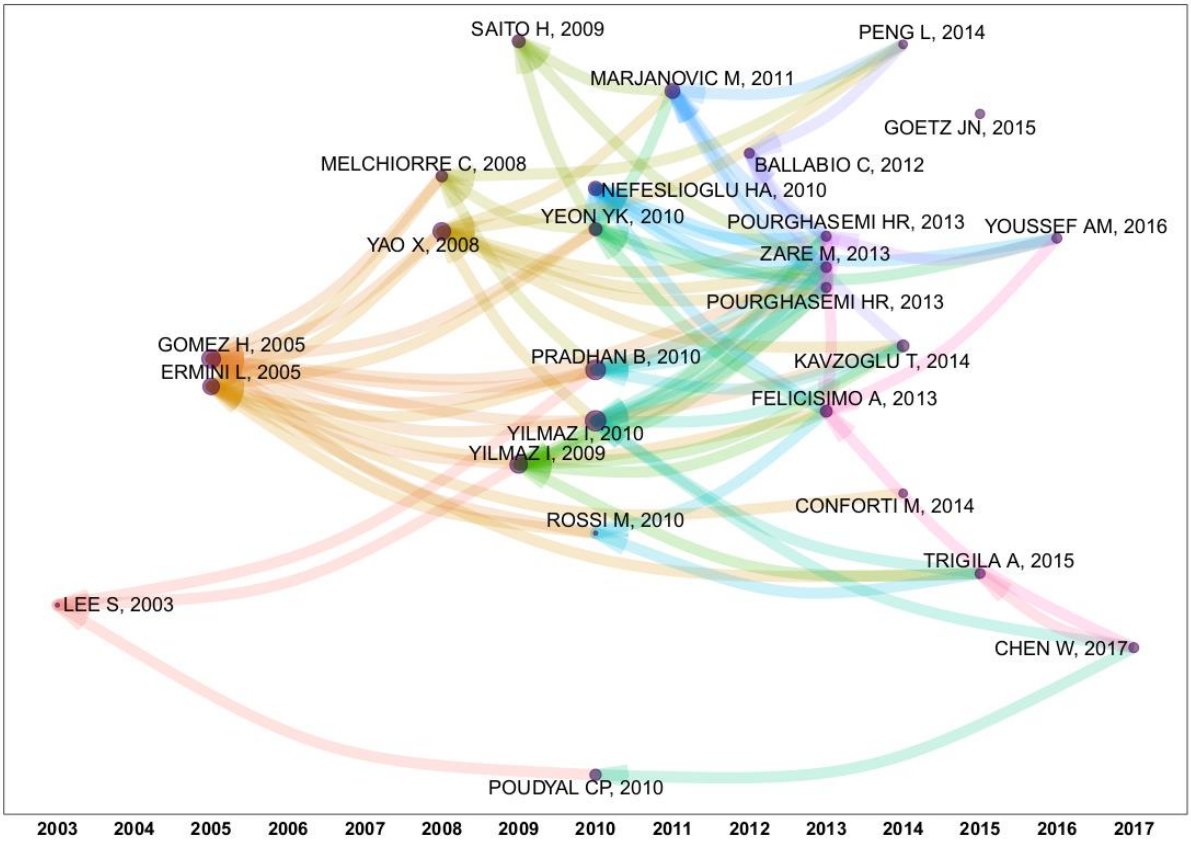
Outra maneira interessante de identificar relações importantes na área de estudo é verificar a rede de citações históricas entre os trabalhos. Conforme exposição da Figura 7, é possível identificar claramente três trabalhos pioneiros que foram citados pelos demais.

O primeiro trabalho, publicado por Lee e outros (2003), explorou a aplicabilidade de Redes Neurais Artificiais para o mapeamento de susceptibilidade a deslizamento de terras. Tal trabalho trouxe avanços significativos ao estabelecer um processo claro para o problema proposto, bem como representou avanços ao utilizar imagens de satélite para a identificação de áreas atingidas por deslizamento. Além disso, o estudo forneceu algumas questões de pesquisa norteadoras para trabalhos futuros. Uma delas relata a impossibilidade de identificar a importância relativa das variáveis ao se utilizar redes neurais. Essa questão foi resolvida ao fazer uso de métodos como o *Random Forest*, que possibilitam a determinação da importância relativa de cada variável.

O segundo trabalho, por sua vez, publicado por Gómez e Kavzoglu (2005), também demostrou o uso de Redes Neurais Artificias para o mapeamento da susceptibilidade de deslizamentos de terra. Tal trabalho mostrou avanço ao utilizar o algoritmo *Multilayer Precepton com Backpropagation*.

O terceiro trabalho identificado, publicado por Ermini; e outros (2005), também fez uso de Redes Neurais Artificias para o mapeamento da susceptibilidade a deslizamentos. O que mostra que a técnica inicial de inteligência artificial para análise do perigo de deslizamentos foi a Rede Neural Artificial.

Figura 7– Rede de citações históricas entre os autores



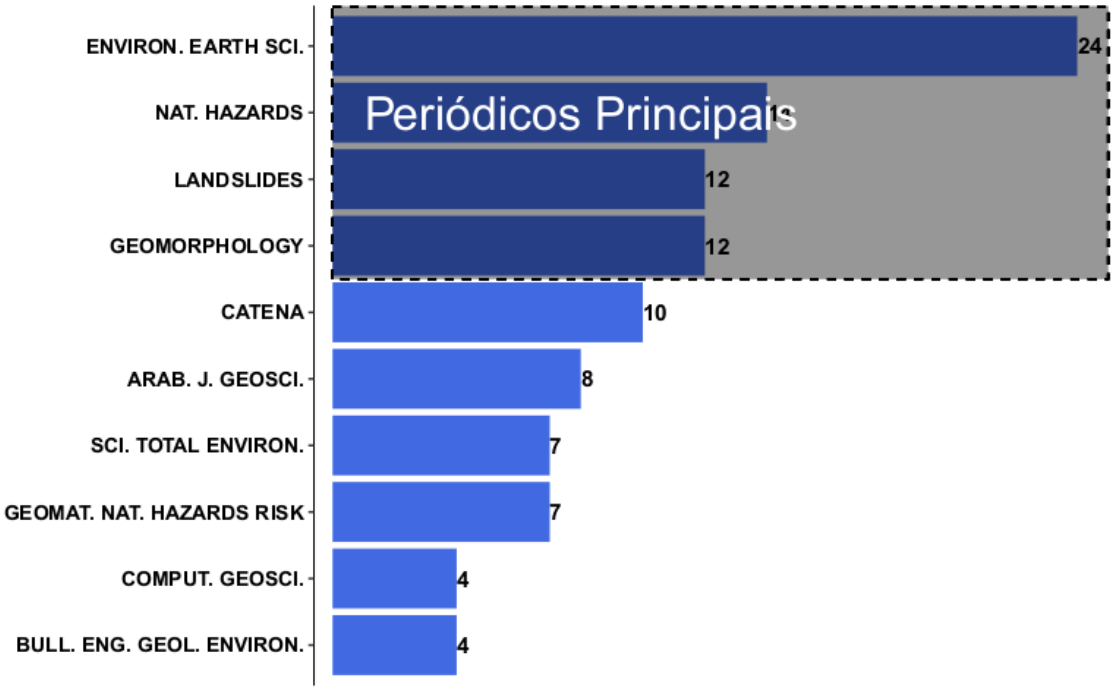
Fonte: O Autor (2019)

A lei de Bradford, surgiu através de pesquisas médicas conduzidas por Hill Bradford. Tal lei permite estimar a relevância de periódicos dentro de áreas específicas do conhecimento, classificando-os em três zonas, cada uma com um terço do total de artigos. A primeira: contém um pequeno número de periódicos altamente produtivos. A segunda: contém um número maior de periódicos menos produtivo. A terceira: contém um número ainda maior de periódicos, reduzindo a produtividade do grupo (MACHADO JUNIOR *et al*., 2016).

Os artigos aqui discutidos estão distribuídos em 38 periódicos diferentes. O Gráfico 5 expõe os dez periódicos com maior número de publicações. Segundo a lei de Bradford, Machado Junior e outros (2016), os periódicos com maior produtividade no tema em questão

são: *Environmental Earth Sciences*, com 24 artigos; *Natural Hazards*, com 14 artigos e; *Landslides e Geomorphology*, com 12 artigos.

Gráfico 5– Citações por periódico



Fonte: O Autor (2019)

A amostra analisada mostrou-se diversificada e significativa. Assim sendo, a próxima etapa consiste em responder as questões de pesquisa previamente estabelecidas. Para isso, cada questão será respondida levando em consideração os dois tipos de desastre analisados no presente trabalho: inundações e deslizamentos de terra. Dessa forma, para cada questão de pesquisa, haverá duas respostas, uma para cada tipo de desastre.

A amostra selecionada possui artigos relacionados a deslizamentos de terra e inundações. Foram identificados 133 artigos para deslizamentos de terra e 22 artigos para inundações. Os artigos categorizados de acordo com cada tipo de desastre estão expostos no Quadro 6.

O trabalho elaborado por Mirzaei e outros (2018) foi considerado como pertencente as duas categorias, por tratar tanto de deslizamentos como de inundações em seu trabalho.

Quadro 6– Artigos classificado por tipo do desastre abordado

Categoria do desastre	Autores
Inundação	(NANDI *et al*., 2016); (TIEN BUI *et al*., 2016a); (TEHRANY *et al*., 2015b); (RAHMATI; POURGHASEMI, 2017); (MIRZAEI *et al*., 2018); (SHAFIZADEH-MOGHADAM *et al*., 2018); (KHOSRAVI *et al*., 2018); (HONG *et al*., 2018); (SAMANTA; PAL; PALSAMANTA, 2018); (AL-ABADI, 2018); (SAMANTA *et al*., 2018); (RAZAVI TERMEH *et al*., 2018); (ZHAO *et al*., 2018); (KOURGIALAS; KARATZAS, 2017); (LEE *et al*., 2017); (LAI *et al*., 2016); (FENG *et al*., 2015); (WANG *et al*., 2015b); (FENG; LIU; GONG, 2015); (JI *et al*., 2013); (PAN *et al*., 2011); (CHANG *et al*., 2010);
Deslizamento de terra	(SHARMA *et al*., 2014); (WANG *et al*., 2016); (ROSSI *et al*., 2010); (GARCÍA-RODRÍGUEZ *et al*., 2008); (GARCÍA-RODRÍGUEZ; MALPICA, 2010); (SU *et al*., 2015); (YOUSSEF *et al*., 2016); (TIEN BUI *et al*., 2012a); (PHAM *et al*., 2016a); (BUI *et al*., 2011); (LOMBARDO *et al*., 2015); (PRADHAN; LEE, 2010); (TIEN BUI *et al*., 2012b); (CHEN *et al*., 2017a); (HONG *et al*., 2015); (KAVZOGLU; SAHIN; COLKESEN, 2014); (GOETZ *et al*., 2015); (TANER SAN, 2014); (YILMAZ, 2010b); (OZDEMIR, 2011); (TIEN BUI *et al*., 2016b); (YAO; THAM; DAI, 2008); (TIEN BUI *et al*., 2016c); (CONFORTI *et al*., 2014); (PHAM *et al*., 2016b); (PENG *et al*., 2014); (FEIZIZADEH *et al*., 2017); (FARAJI SABOKBAR; SHADMAN ROODPOSHTI; TAZIK, 2014); (TRIGILA *et al*., 2015); (ZARE *et al*., 2013); (XU *et al*., 2013); (GOKCEOGLU *et al*., 2010); (SANGCHINI *et al*., 2016); (FENG *et al*., 2016); (SUJATHA *et al*., 2013); (NEFESLIOGLU *et al*., 2011); (ERCANOGLU; TEMIZ, 2011); (MELCHIORRE *et al*., 2011); (YI-TING *et al*., 2015); (TEHRANY; PRADHAN; JEBUR, 2015); (POURGHASEMI; MORADI; FATEMI AGHDA, 2013); (YILMAZ, 2009a); (PRADHAN; PUTRA, 2013); (KINCAL; AKGUN; KOCA, 2009); (INTARAWICHIAN; DASANANDA, 2011); (YILMAZ, 2009b); (PHAM *et al*., 2017); (YILMAZ, 2010a); (POLYKRETIS; FERENTINOU; CHALKIAS, 2014); (TIEN BUI *et al*., 2017); (BALLABIO; STERLACCHINI, 2012); (CHEN *et al*., 2014); (ARNONE *et al*., 2014); (DAHAL, 2014); (POURGHASEMI; KERLE, 2016); (COSTANZO *et al*., 2014); (PARK; LEE, 2014); (HONG; POURGHASEMI; POURTAGHI, 2016); (HONG *et al*., 2016); (AKGUN; KINCAL; PRADHAN, 2012); (TSANGARATOS; ILIA, 2016); (REGMI *et al*., 2014); (SOLAIMANI; MOUSAVI; KAVIAN, 2013); (WANG *et al*., 2015a); (ALTHUWAYNEE *et al*., 2014); (WANG; SAWADA; MORIGUCHI, 2013); (FELICÍSIMO *et al*., 2013); (DEMIR *et al*., 2013); (CHEN *et al*., 2016); (GURI; CHAMPATI RAY; PATEL, 2015); (KAYASTHA; DHITAL; DE SMEDT, 2013); (HONG *et al*., 2017a); (SEGONI *et al*., 2015); (HUANG; ZHAO, 2018); (BALAMURUGAN; RAMESH; TOUTHANG, 2016); (CHEN *et al*., 2018a); (ZHU *et al*., 2018); (SUN *et al*., 2018); (POURGHASEMI *et al*., 2018); (BUI *et al*., 2018); (BORNAETXEA *et al*., 2018); (LEE; LEE; LEE, 2018); (ADINEH *et al*., 2018); (POLYKRETIS; CHALKIAS, 2018); (WANG *et al*., 2018); (MIRZAEI *et al*., 2018); (MERGHADI; ABDERRAHMANE; TIEN BUI, 2018); (HOANG; TIEN BUI, 2018); (CHEN *et al*., 2018b); (CHEN; POURGHASEMI; NAGHIBI, 2018a); (CHEN; POURGHASEMI; NAGHIBI, 2018b); (CHEN *et al*., 2018c); (MONDAL; MANDAL, 2018); (POURGHASEMI; RAHMATI, 2018); (PHAM, 2018); (LEE *et al*., 2018); (PHAM; PRAKASH; TIEN BUI, 2018); (PHAM; TIEN BUI; PRAKASH, 2018); (BEHNIA; BLAIS-STEVENS, 2018); (KIM *et al*., 2018); (KALANTAR *et al*., 2018); (LIU; MIAO, 2018); (ADA; SAN, 2018); (ARABAMERI; POURGHASEMI; YAMANI, 2017); (PHAM; TIEN BUI; PRAKASH, 2017); (CHEN *et al*., 2017b); (SHRESTHA; KANG; SUWAL, 2017); (HONG *et al*., 2017b); (POURGHASEMI; ROSSI, 2017); (WANG *et al*., 2017); (TSANGARATOS *et al*., 2017); (ZHANG *et al*., 2017b); (NGUYEN *et al*., 2017); (XIE *et al*., 2017); (CHEN *et al*., 2017c); (LEE *et al*., 2015); (SHAHABI; HASHIM; AHMAD, 2015); (TSAI *et al*., 2013); (RAMAKRISHNAN *et al*., 2013); (LEE; HWANG; PARK, 2013); (RAMANI SUJATHA; KUMARAVEL; RAJAMANICKAM G, 2012); (TIEN BUI *et al*., 2012c); (LEE; OH, 2014); (RAMANI; PITCHAIMANI; GNANAMANICKAM, 2011); (MOUSAVI *et al*., 2011); (YEON; HAN; RYU, 2010); (POUDYAL *et al*., 2010); (PRABU; RAMAKRISHNAN, 2009); (LEE, 2007); (GÓMEZ; KAVZOGLU, 2005); (ERMINI; CATANI; CASAGLI, 2005); (LEE *et al*., 2003)

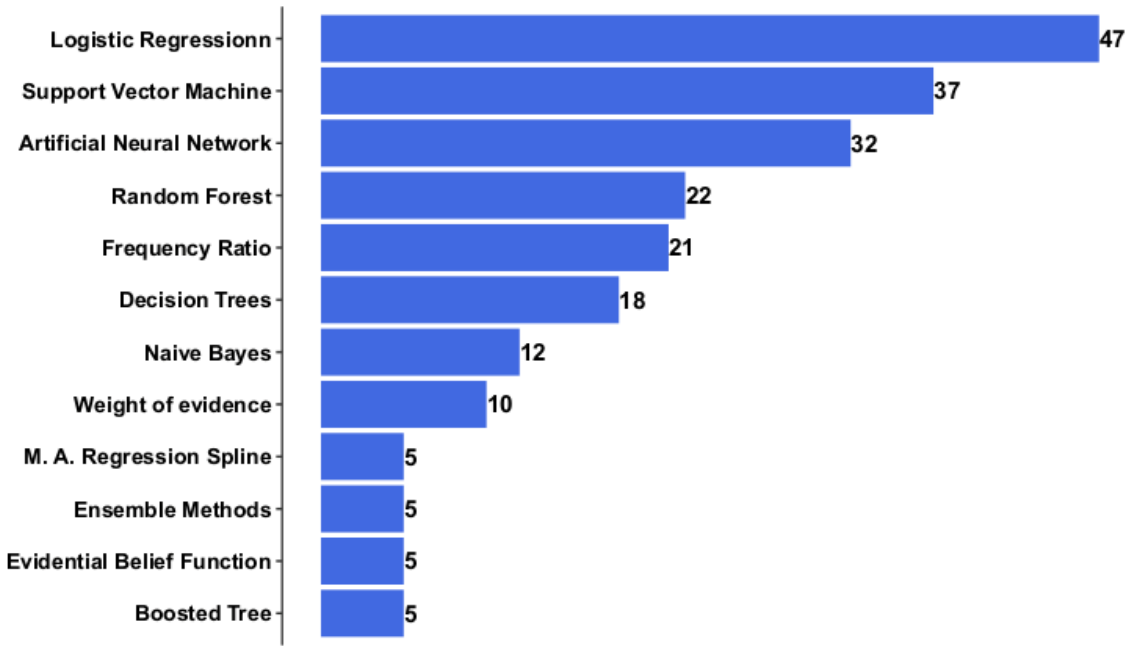
Fonte: O Autor (2019)

4.2.2 Quais os métodos de aprendizado de máquina são mais utilizados para o mapeamento do perigo de enchentes e inundações?

Foram identificados 61 modelos diferentes no presente estudo. Para mapeamento do perigo de deslizamentos foram utilizados 54 desses modelos. Já para o mapeamento do perigo de inundação 23 desses modelos foram usados.

Dentre os 54 modelos utilizados para mapear o perigo de deslizamento, *Logistic Regression* foi o mais utilizado, seguido por SVM, ANN, *Random Forest* e *Frequency Ratio*. Para fins de simplificação visual, o Gráfico 6 mostra apenas modelos que foram utilizados pelo menos 5 vezes.

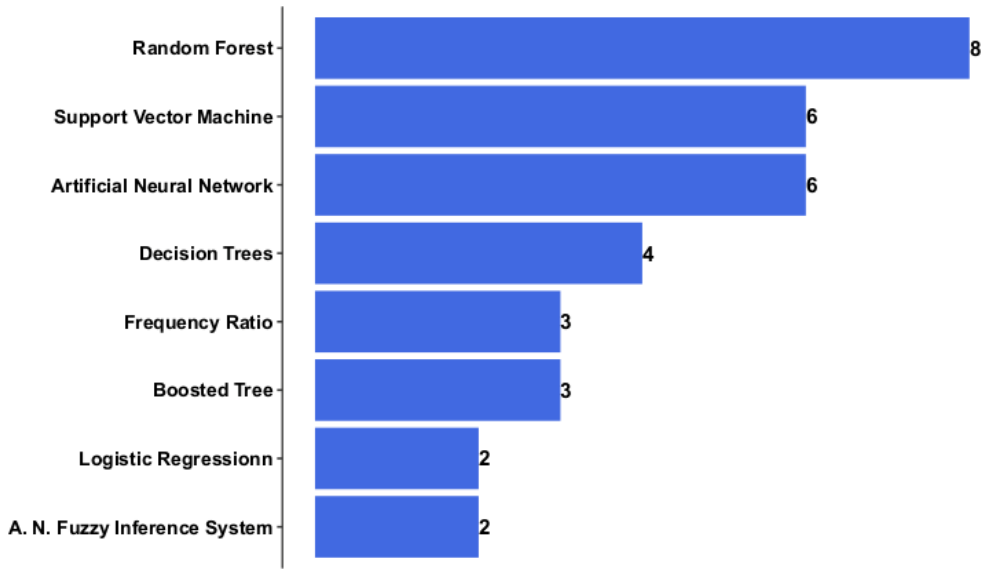
Gráfico 6 – Modelos mais utilizados para mapeamento de perigo de deslizamentos



Fonte: O Autor (2019)

Para inundações, o modelo mais utilizado foi o *Random Forest*, seguido por SVM, ANN, *Decision Trees* e *Frequency Ratio*. O Gráfico 7 exibe apenas os modelos que forma usados pelo menos duas vezes.

Gráfico 7– Modelos mais utilizados para mapeamento de perigo de inundação



Fonte: O Autor (2019)

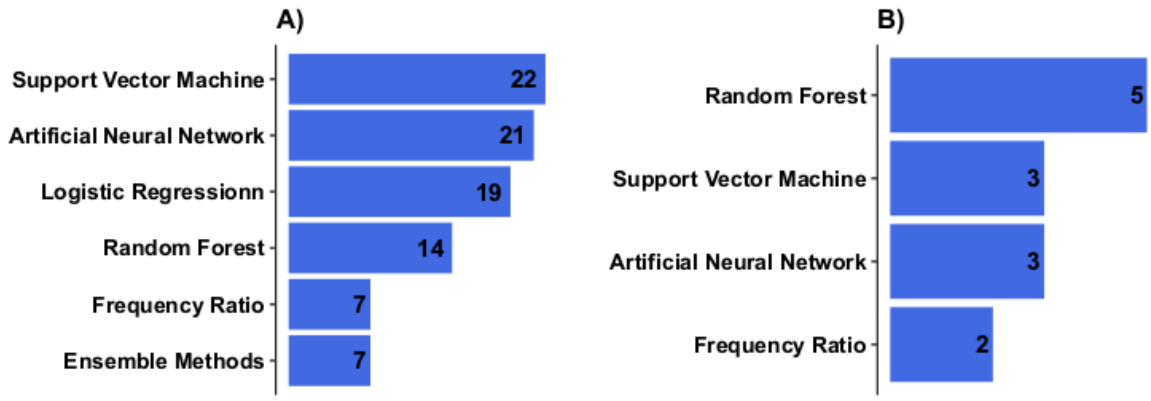
Ao analisar os cinco modelos mais utilizados, é possível perceber apenas leves diferenças entre os resultados. O primeiro, o modelo *Logistic Regression*, o mais utilizado para mapeamento do perigo de deslizamentos, não aparece entre os cinco primeiros para o mapeamento do perigo de inundações. Em segundo lugar, o modelo *Decision Trees*, quarto colocado para mapeamento de perigo de inundações, não aparece entre os cinco mais utilizados para deslizamento. Por fim, há diferenças nas posições que cada modelo assume para cada tipo de desastre.

Nos estudos analisados, frequentemente um único artigo fez uso de mais de um algoritmo, objetivando a comparação entre eles e a escolha do que apresentou o melhor desempenho. Dessa forma, foi possível catalogar quais modelos apresentaram melhor desempenho para cada artigo. Tal informação está exposta no Gráfico 8.

Novamente é possível perceber a presença dos modelos discutidos nos parágrafos anteriores. As primeiras colocações foram ocupadas por modelos conhecidos e amplamente utilizados nas tarefas de mapeamento dos desastres hidrológicos.

Foram aplicados filtros para produzir o Gráfico 8. Para deslizamentos só foram considerados modelos que foram melhores que os demais no mínimo 5 vezes. Já para inundações, foram considerados modelos que foram melhores que os demais no mínimo 2 vezes.

Gráfico 8– Modelos com melhor desempenho a) para deslizamento; b) para inundação



Fonte: O Autor (2019)

Os modelos usados em deslizamentos de terra mostram um grupo bem definido de algoritmos superiores em performance. Esse grupo é formado por *SVM*, *ANN*, *Logistic Regression* e *Random Forest*. Esse resultado indica que tais modelos são amplamente utilizando e confiáveis para tal tarefa.

Tendo em vista as análises realizadas até o momento, foi possível estabelecer um conjunto de modelos mais utilizados e com performance superior. Dentro desse grupo estão: *ANN*; *SVM*; *Random Forest*; *Logistic Regression* e; *Frequency Ratio*. Os modelos incluídos nesse grupo foram escolhidos seguindo os critérios frequência de utilização de desempenho nas tarefas.

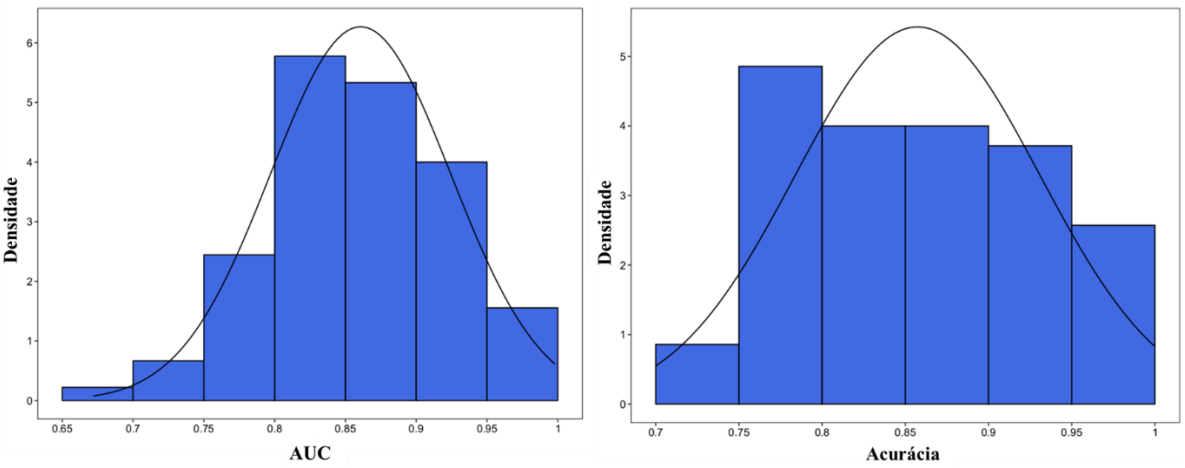
4.2.3 Como estão distribuídos os valores da avaliação de performance dos métodos de aprendizado de máquina para o mapeamento do perigo de enchentes e inundações?

Como critério de avaliação dos modelos foram usadas duas medidas amplamente utilizadas. A primeira e mais comum entre os artigos selecionados foi a AUC, apresentada e discutida no capítulo 2, que foi usada em 90 dos artigos selecionados. Já a segunda, em menor número nos artigos selecionados foi a acurácia, presente em 70 dos artigos. Nove dos artigos analisados utilizaram tanto AUC como acurácia.

Vale ressaltar que trabalhos como o de Pan e outros (2011) utilizam a abordagem de regressão, o que inviabiliza a utilização das medidas adotadas como padrão nesse estudo. Dentro da amostra, cinco artigos foram excluídos da análise de desempenho devido ao fato explicado anteriormente. Apesar de configurar uma limitação, tal quantidade não é suficiente para inviabilizar a análise.

Com o intuito de estabelecer um critério de comparação para estudos futuros, foram realizados testes de aderência com distribuições conhecidas. O teste de Shapiro-Wilk retornou valor-p de 0,3022, dessa forma, não podemos rejeitar a hipótese de normalidade dos dados, com média 0,8607 e desvio padrão de 0,0636. Já para a métrica acurácia a distribuição normal não obteve bom ajuste. O teste de *Shapiro-Wilk* retornou valor-p de 0,02731, fornecendo evidência suficiente para rejeitar a hipótese de normalidade dos dados. Com média de 0,8574 e desvio padrão de 0,0735. No Gráfico 9 está exposto a distribuição dos dados e a distribuição normal usada para o teste de aderência.

Gráfico 9– Desempenho dos modelos segundo a métrica acurácia



Fonte: O Autor (2019)

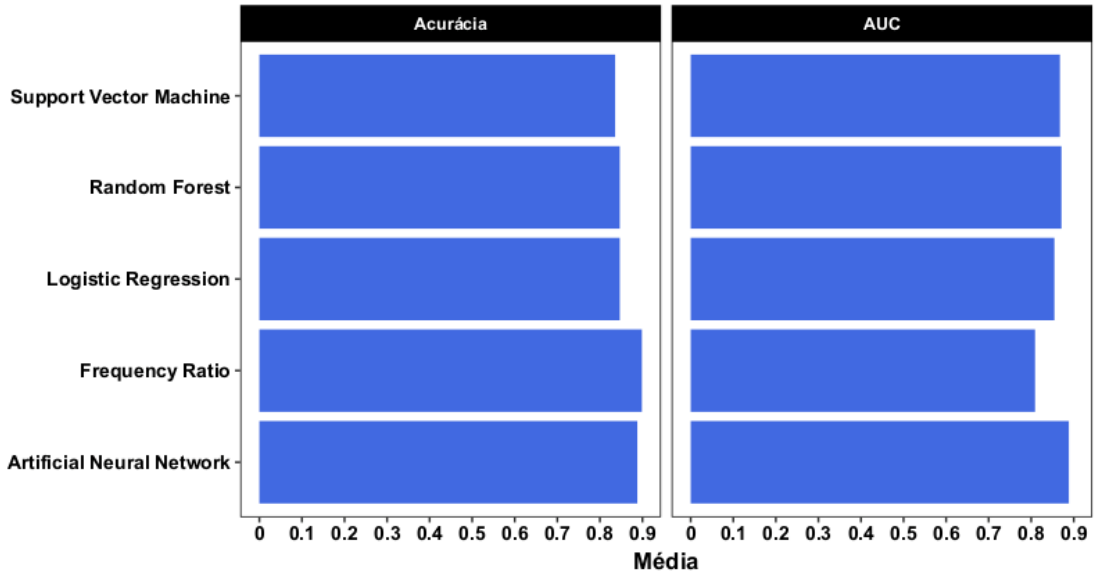
Ao analisar o desempenho dos modelos mais utilizados para a tarefas de mapeamento de perigo de deslizamento e inundação, não é possível perceber diferença significativa, como explicitado na Tabela 2 e Gráfico 10.

Tabela 2– Padrão de desempenho dos modelos mais utilizados

	AUC			Acurácia		
Modelo	Observações	Media	Desvio padrão	Observações	Média	Desvio padrão
Artificial Neural Network	14	0,8884	0,0513	7	0,8880	0,0650
Frequency Ratio	6	0,8090	0,0400	3	0,8987	0,0688
Logistic Regression	12	0,8544	0,0520	9	0,8468	0,0836
Random Forest	5	0,8710	0,0965	15	0,8467	0,0727
Support Vector Machine	20	0,8674	0,0714	9	0,8359	0,0736

Fonte: O Autor (2019)

Gráfico 10– Desempenho dos modelos mais utilizados



Fonte: O Autor (2019)

A diferença entre o desempenho dos modelos apresentados tanto no Gráfico 10 quanto na Tabela 2 não apresentaram diferença significativa. Esse fato mostra que qualquer um desses modelos pode ser utilizado com segurança, pois são capazes de entregar desempenho equivalentes quando comparados com os demais.

Por fim, é possível estabelecer um padrão para comparação com trabalhos futuros. Além dos parâmetros das distribuições ajustadas, será fornecido os quartis, para possibilitar a localização não paramétrica em termos de desempenho, seguindo as duas métricas aqui adotadas. A Tabela 3 resume tais informações.

Tabela 3 – Padrão de desempenho dos modelos

	Acurácia	**AUC**
Distribuição	Não normal	Normal
Mediana	0,8550	0,8530
Média	0,8574	0,8607
Desvio padrão	0,0735	0,0636
Valor-p para o teste de *Shapiro-Wilk*	0,02731	0,3022
Quartil 1 (0 – 25%)	[0,7100, 0,7926)	[0,6721, 0,8185)
Quartil 2 (25% – 50%)	[0,7926, 0,8550)	[0,8185, 0,8525)
Quartil 3 (50% – 75%)	[0,8550, 0,9195)	[0,8525, 0,9092)
Quartil 4 (75% – 100%)	[0,9195, 1]	[0,9092, 1]

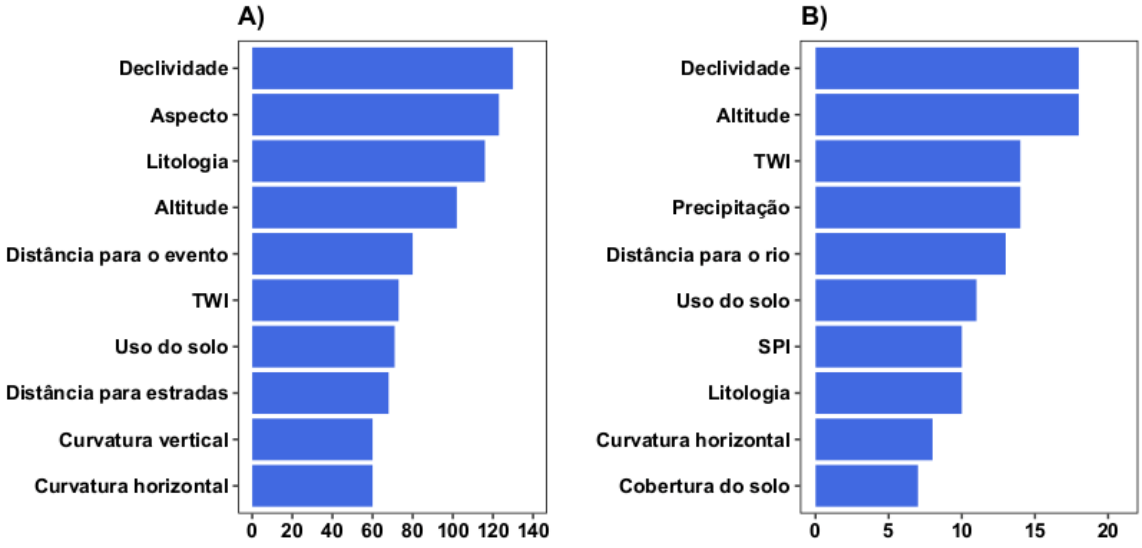
Fonte: O Autor (2019)

4.2.4 Quais as variáveis condicionantes mais utilizadas nos métodos de aprendizado de máquina para o mapeamento do perigo de enchentes e inundações?

Foram identificadas 94 variáveis condicionantes diferentes. Dentre essas, algumas delas mostraram-se com alta frequência de utilização, já outras, foram utilizadas apenas em algumas situações específicas. Também foi evidenciado uma leve variação entre as variáveis utilizadas para deslizamentos e inundações.

O Gráfico 11 mostra as 10 variáveis condicionantes mais utilizadas. Como é possível observar, há diferenças entre os dois tipos de desastres. Variáveis como distância para o rio, distância para o evento, Distância para as rodovias estão presentes entre as mais usadas apenas para um tipo de desastre. Isso mostra, como esperado, que para cada tipo de desastre variáveis condicionantes diferentes devem ser utilizadas.

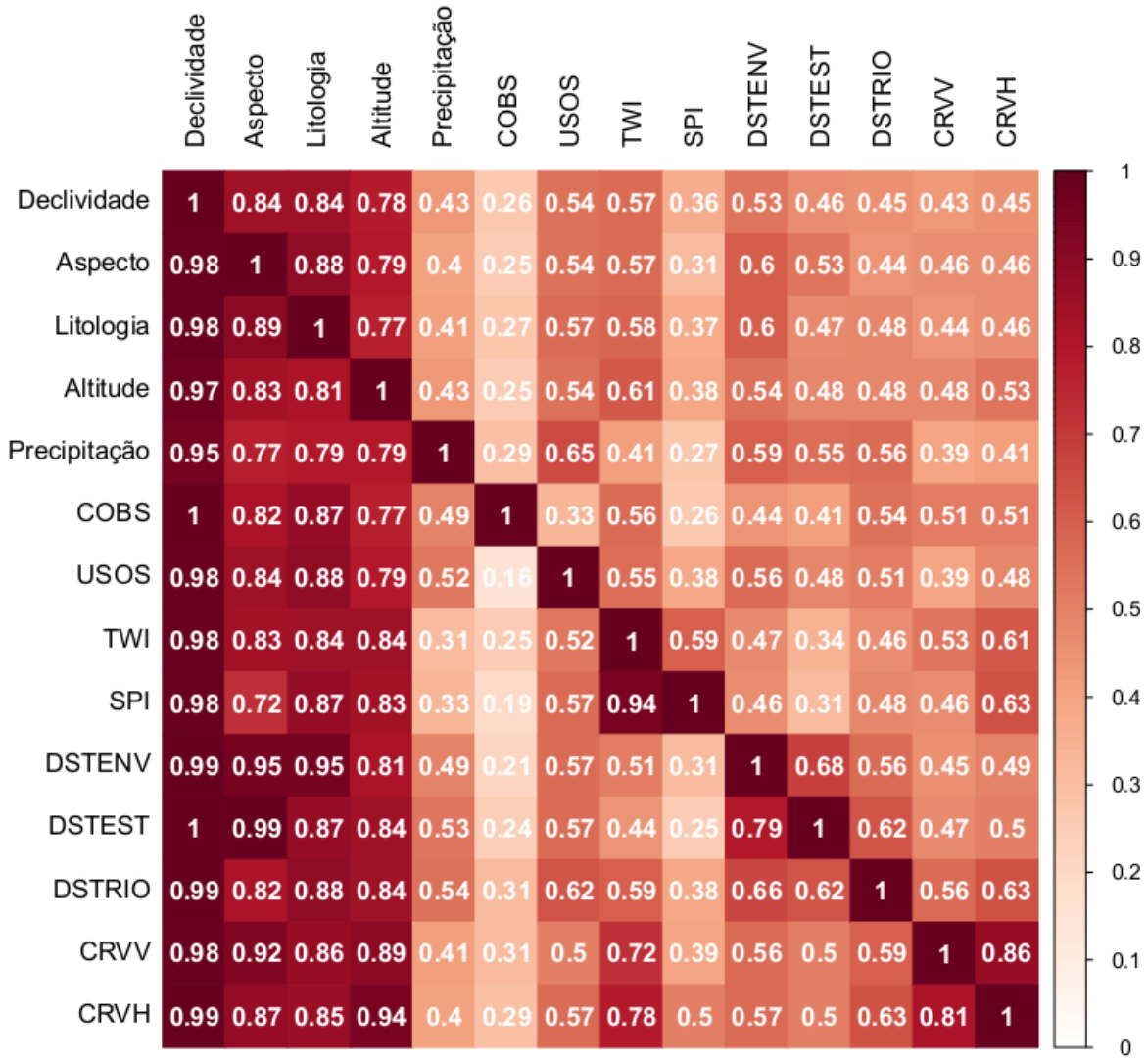
Gráfico 11– Variáveis condicionantes para A) deslizamentos; B) inundações



Fonte: O Autor (2019)

Tais variáveis representam um grupo significativo e comumente usados para a tarefa em questão. Além disso, como mostra a Figura 8, existe um grupo com frequência de co-ocorrência relativa alta. Por exemplo, quando a variável declividade foi utilizada, em 84% das vezes as variáveis aspecto e litologia também foram utilizadas. Além disso, o grupo formado por declividade, aspecto, litologia e altitude possuem alta frequência de co-ocorrência para todas as variáveis analisadas, indicando que tal grupo é frequentemente usado independentemente das demais variáveis consideradas. O significado de tais variáveis são explicados no capítulo 5.

Figura 8– Frequência relativa entre as variáveis



	Declividade	Aspecto	Litologia	Altitude	Precipitação	COBS	USOS	TWI	SPI	DSTENV	DSTEST	DSTRIO	CRVV	CRVH
Declividade	1	0.84	0.84	0.78	0.43	0.26	0.54	0.57	0.36	0.53	0.46	0.45	0.43	0.45
Aspecto	0.98	1	0.88	0.79	0.4	0.25	0.54	0.57	0.31	0.6	0.53	0.44	0.46	0.46
Litologia	0.98	0.89	1	0.77	0.41	0.27	0.57	0.58	0.37	0.6	0.47	0.48	0.44	0.46
Altitude	0.97	0.83	0.81	1	0.43	0.25	0.54	0.61	0.38	0.54	0.48	0.48	0.48	0.53
Precipitação	0.95	0.77	0.79	0.79	1	0.29	0.65	0.41	0.27	0.59	0.55	0.56	0.39	0.41
COBS	1	0.82	0.87	0.77	0.49	1	0.33	0.56	0.26	0.44	0.41	0.54	0.51	0.51
USOS	0.98	0.84	0.88	0.79	0.52	0.16	1	0.55	0.38	0.56	0.48	0.51	0.39	0.48
TWI	0.98	0.83	0.84	0.84	0.31	0.25	0.52	1	0.59	0.47	0.34	0.46	0.53	0.61
SPI	0.98	0.72	0.87	0.83	0.33	0.19	0.57	0.94	1	0.46	0.31	0.48	0.46	0.63
DSTENV	0.99	0.95	0.95	0.81	0.49	0.21	0.57	0.51	0.31	1	0.68	0.56	0.45	0.49
DSTEST	1	0.99	0.87	0.84	0.53	0.24	0.57	0.44	0.25	0.79	1	0.62	0.47	0.5
DSTRIO	0.99	0.82	0.88	0.84	0.54	0.31	0.62	0.59	0.38	0.66	0.62	1	0.56	0.63
CRVV	0.98	0.92	0.86	0.89	0.41	0.31	0.5	0.72	0.39	0.56	0.5	0.59	1	0.86
CRVH	0.99	0.87	0.85	0.94	0.4	0.29	0.57	0.78	0.5	0.57	0.5	0.63	0.81	1

COBS: cobertura do solo; USOS: uso do solo; TWI: Topographic wetness index; SPI: stream power index; DSTEST: Distância para as rodovias; DSTRIO: distância para o rio; CRVV: curvatura vertical; CRVH; curvatura horizontal.

Fonte: O Autor (2019)

Entretanto, várias outras variáveis foram utilizadas para a tarefa em questão. A utilização de tais variáveis deve levar em conta o contexto e a disponibilidade de dados. A Tabela 4 exibe as variáveis que foram utilizadas pelo menos 2 vezes para um dos tipos de desastre. É possível notar que devido à variedade de variáveis condicionantes utilizadas, os modelos ajustados sofrem influência direta do contexto no qual estão sendo modelados. Isso põe em discussão uma importante questão, pois cada modelo é pensado para atender a uma necessidade específica de uma localidade, incluindo em sua estrutura fatores específicos de cada área, dessa forma, a

utilização dos modelos em outras áreas deve ser analisada com cautela, levando em consideração as diferenças de cada área de estudo.

Tabela 4 – Variáveis utilizadas nos modelos analisados

Variável	Deslizamento	Inundação	Variável	Deslizamento	Inundação
Declividade	130	18	**Densidade da drenagem**	13	3
Aspecto	123	4	**Comprimento do ângulo**	12	1
Litologia	116	10	**Densidade da floresta**	11	0
Altitude	102	18	**TRI**	10	1
TWI	73	14	**Diâmetro da floresta**	10	0
Uso do solo	71	11	**Intemperismo**	8	2
Distância para o evento	80	0	**Idade da floresta**	9	0
Distância para as rodovias	68	0	**Densidade de estradas**	8	1
Curvatura horizontal	60	8	**Profundidade do solo**	8	0
Distância para o rio	55	13	**Densidade lineamento**	7	0
Precipitação	52	14	**Acumulação de fluxo**	6	1
Curvatura vertical	60	4	**Densidade de eventos**	6	0
SPI	44	10	**Índice de convergência**	6	0
NDVI	37	5	**Amplitude**	6	0
Curvatura do terreno	38	2	**Aceleração do movimento do solo**	4	0
Cobertura do solo	32	7	**Convexidade**	4	0
Distância para a drenagem	29	0	**Número de escoamento**	1	3
Tipo do solo	24	4	**Distância para as placas tectônicas**	3	0
Textura do solo	17	6	**TPI**	3	0
Relevo	19	3	**Altitude relativa**	2	1
STI	20	1	**NDWI**	2	1
Drenagem do solo	16	3	**Permeabilidade**	2	1
Tipo da floresta	15	3	**Densidade populacional**	1	2
Densidade do rio	13	5	**Frequência de tempestade**	1	2
Distância para o lineamento	17	0			

Fonte: O Autor (2019)

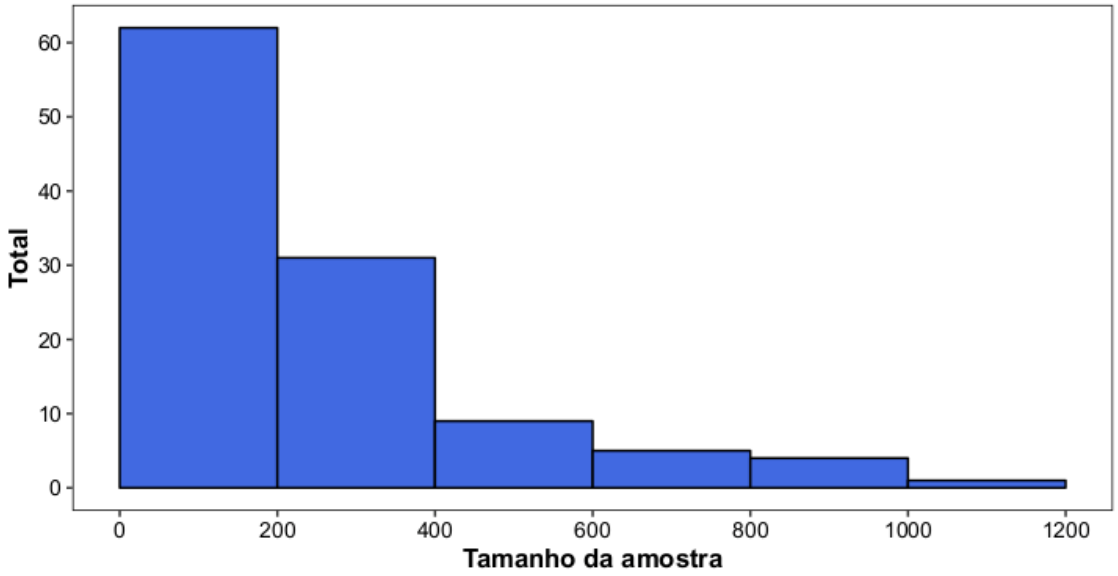
4.2.5 Como estão distribuídos o tamanho das amostras utilizadas nos métodos de aprendizado de máquina para o mapeamento do perigo de enchentes e inundações?

Os tamanhos das amostras utilizados variam muito entre os tipos de desastres e até mesmo dentro das categorias. Assim sendo, será realizado o tratamento para identificar *outliers* pela métrica de Tukey (1977), que considera *outliers* valores menores que $Q_1 – 1,5(Q_3 – Q_1)$ e valores maiores que $Q_3 + 1,5(Q_3 - Q_1)$, onde Q_i representa o valor do quartil i, os quais serão caracterizados e posteriormente retirados da amostra.

Dentre os 154 trabalhos analisados, 22 deles foram identificados como fazendo uso de amostras muito grandes. Para os trabalhos que abordaram deslizamentos, 18 deles foram considerados *outliers*, com valores acima de 1156 deslizamentos. Já para os trabalhos que abordaram o desastre de inundação, apenas 4 deles foram identificados como *outliers*, com valores acima de 3540 eventos.

Como mostrado no Gráfico 12, a maior concentração está em tamanhos de amostra entre 46 e 400 eventos analisados, com mais de noventa eventos. Os valores acima de 1156 foram considerados *outliers* e excluídos do Gráfico 12. O maior tamanho de amostra analisado foi de 48007.

Gráfico 12– Tamanho da amostra para estudos de deslizamento

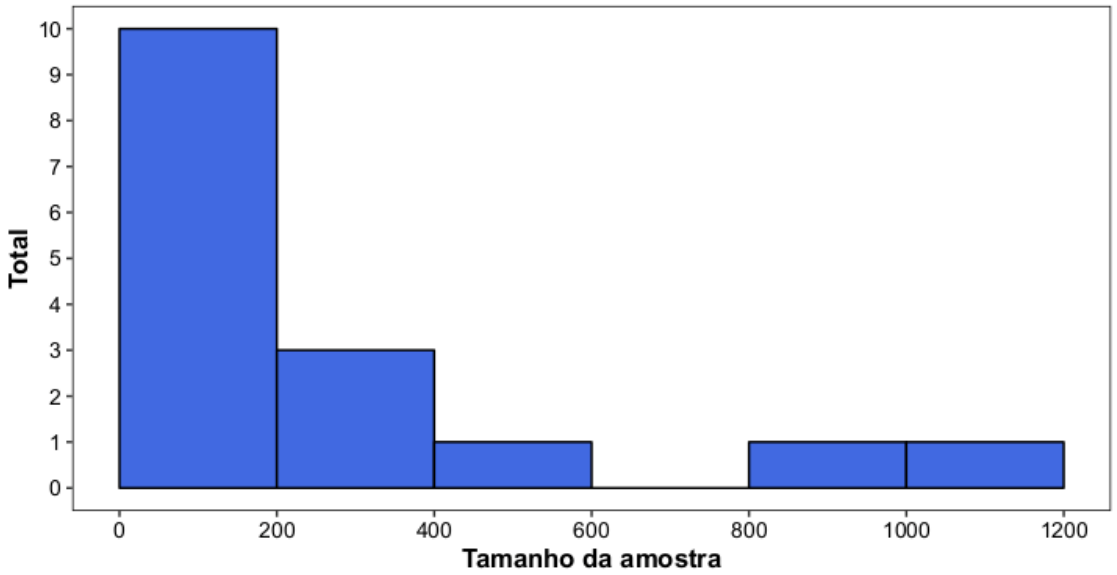


Fonte: O Autor (2019)

Com relação ao tipo de evento inundação a maior frequência de tamanho de amostra está entre 9 e 400 eventos analisados, como mostra o Gráfico 13. Os valores acima de 3540 foram

considerados *outliers*. O maior tamanho de amostra analisado nos trabalhos de inundação foi de 27108 eventos.

Gráfico 13– Tamanho da amostra para estudos de deslizamento



Fonte: O Autor (2019)

Com o objetivo de fornecer um padrão de comparação, os quartis das amostras estão expostos na Tabela 5.

Tabela 5 – Quartis da distribuição do tamanho das amostras analisadas

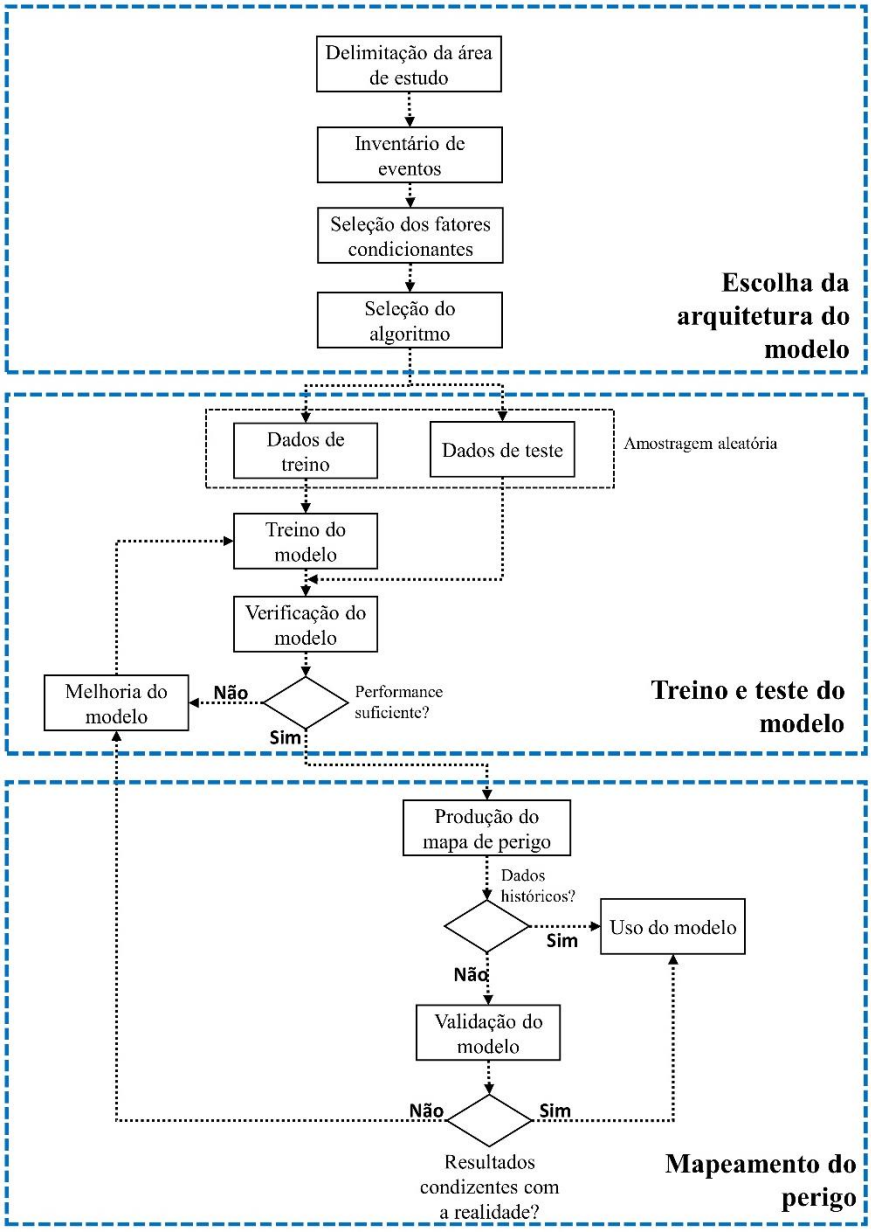
Quartil	Inundações	Deslizamentos
0 – 25%	[46; 75)	[9;99)
25% – 50%	[75; 152)	[99; 175)
50% – 75%	[152; 236)	[175; 328)
75% – 100	[236; 1160)	[328; 1102)

Fonte: O Autor (2019)

4.3 Processo de análise de perigo de inundação e deslizamento com modelos de Aprendizado de Máquina

Uma visão geral do processo proposto nesse trabalho para o mapeamento de perigo de deslizamentos e inundações com algoritmos de aprendizado de máquina (FLSM-ML) é mostrada no Fluxograma 7. Consiste principalmente de três fases. A primeira para escolher a arquitetura do modelo. A segunda para treinar e testar a performance do modelo. Por fim, a terceira para elaborar o mapa de susceptibilidade a perigo.

Fluxograma 7 – Processo metodológico utilizado no estudo



Fonte: esta pesquisa (2019)

A delimitação da área de estudo deve ser feita de maneira a deixar claro qual a escala do estudo, informar quais os limites geográficos, além de fornecer informações sobre as características morfológicas e geológicas da área. Nessa etapa, é importante utilizar ferramentas de visualização para auxiliar o entendimento da informação, como fazem Ramani e outro. (2012) e Sujatha e outros (2013), no qual utilizaram uma abordagem hierárquica, partindo da macrorregião até chegar a região objeto do estudo. Além disso, informações gerais sobre fatores morfológicos e geológicos devem ser fornecidos (CAN *et al*., 2019; COSTACHE, 2019b; LAI *et al*., 2016b; MA *et al*., 2019b; PHAM; PRAKASH, 2019a).

O inventário de eventos deve conter informações sobre onde e quando os eventos ocorreram, o tipo, o tamanho e a extensão do desastre, bem como informações sobre a relação entre a distribuição espacial dos eventos e as diferentes variáveis condicionantes (CHEN *et al*., 2017b; COSTACHE, 2019a; GODT *et al*., 2008). Atenção especial deve ser dada a essa etapa, pois tem efeito direto na qualidade do estudo (CAN *et al*., 2019). O inventário de eventos pode ser construído através de registros históricos, interpretação de imagens e informações obtidas através de sensoriamento remoto, ou ainda, observações em campo (CHEN *et al*., 2018a; REGMI *et al*., 2014; TIAN *et al*., 2019; YI-TING *et al*., 2015). As informações fornecidas na seção 4.2.5 a respeito do tamanho da amostra devem ser utilizadas para fins de consulta às boas práticas.

A seleção das variáveis condicionantes varia bastante entre áreas de estudo de acordo com as características específicas de cada local (TEHRANY; PRADHAN; JEBUR, 2013). A escolha das variáveis podem variar ainda segundo o algoritmo que será utilizado, pois alguns deles para serem utilizados são necessárias suposições a respeito da relação entre as variáveis (CHEN *et al*., 2019d; GAN *et al*., 2012; HONG *et al*., 2015). No presente estudo foi levantada uma extensa lista de variáveis condicionantes que já foram utilizadas para a tarefa em questão, discutidas na seção 4.2.4, tal conjunto de variáveis pode ser usado para guiar os pesquisadores na escolha das variáveis que melhor representam o problema a ser estudado.

A escolha do método adequado para o estudo pode variar segundo alguns critérios, como por exemplo, as informações esperadas, a quantidade de dados disponíveis, capacidade de processamento disponível, entre outros. Na seção 4.2.2 foi discutido um conjunto de métodos que são frequentemente utilizados e produzem bons resultados. Segundo os resultados dessa pesquisa, os algoritmos SVM, ANN, *Random Forest*, *Logistic Regression* e *Frquency Ration* são amplamente usados e capazes de produzem bons resultados. Uma boa estratégia para selecionar qual algoritmo deve ser usado, consiste em realizar testes de performance com um conjunto de algoritmos pré-estabelecidos e verificar qual obteve melhor resultado em termos de acurácia (KUTLUG SAHIN; COLKESEN, 2019; LI *et al*., 2019; MA *et al*., 2019a; MERGHADI; ABDERRAHMANE; TIEN BUI, 2018; SHAO *et al*., 2019; TRIGILA *et al*., 2015).

Uma vez que a arquitetura do modelo foi definida, os dados devem ser particionados em treino e teste. Para isso, geralmente se utiliza a proporção 80% para treino e 20% para teste (ABEDINI *et al*., 2019b; DEMIR *et al*., 2013; XIE *et al*., 2017), porém, não há um consenso sobre tal proporção. O conjunto de dados deve ter eventos positivos bem como eventos negativos, ou seja, amostras nas quais ocorreu um desastre e amostras nas quais não ocorreu

nenhum desastre. A esse respeito, Hong e outros (2019) realizam um estudo sobre a proporção ideal entre dados positivos e dados negativos, chegaram à conclusão que quando a área de amostragem disponível para a amostragem de dados negativos é muito grande, a melhor proporção é 1:1. É recomendado ler o trabalho completo para verificar a proporção ideal em outras situações.

A etapa de verificação do modelo pode ser realizada utilizando métricas de validação cruzada ou AUC. Vale ressaltar que existem outras métricas que podem ser utilizadas a depender das características do problema, como por exemplo, nas situações de dados desbalanceados Castro e Braga (2012). Caso a performance não esteja satisfatória, o modelo deve ser ajustado e reavaliado. A Tabela 2 fornece parâmetros detalhados para comparação. Segundo os resultados obtidos nessa pesquisa, 75% dos modelos que são publicados, possuem performance maior que 0,8, tanto segundo a métrica de acurácia como AUC.

Caso a performance esteja dentro dos padrões esperados, o mapa de perigo pode ser elaborado. Para isso é recomendado a utilização de GIS para produzir representações visuais espaciais do perigo do desastre. A maioria dos estudos aqui analisados utilizaram alguma solução GIS, como Tehrany e outros (2015a), Sciarra e outros (2017), Chen e outros (2018a), entre outros.

Um outro ponto importante são os casos nos quais os dados utilizados para treino do modelo não são provenientes de dados históricos ou observações em campo. Nessas situações é recomendado realizar a validação do modelo, ou seja, verificar se os resultados do modelo condizem com os fatos. Essa tarefa pode ser realizada através de comparação com notícias, junto a autoridades locais ou através da comparação com dados seguros.

Finalmente, caso o modelo possua desempenho adequado, o mesmo pode ser utilizado para os fins previamente planejados.

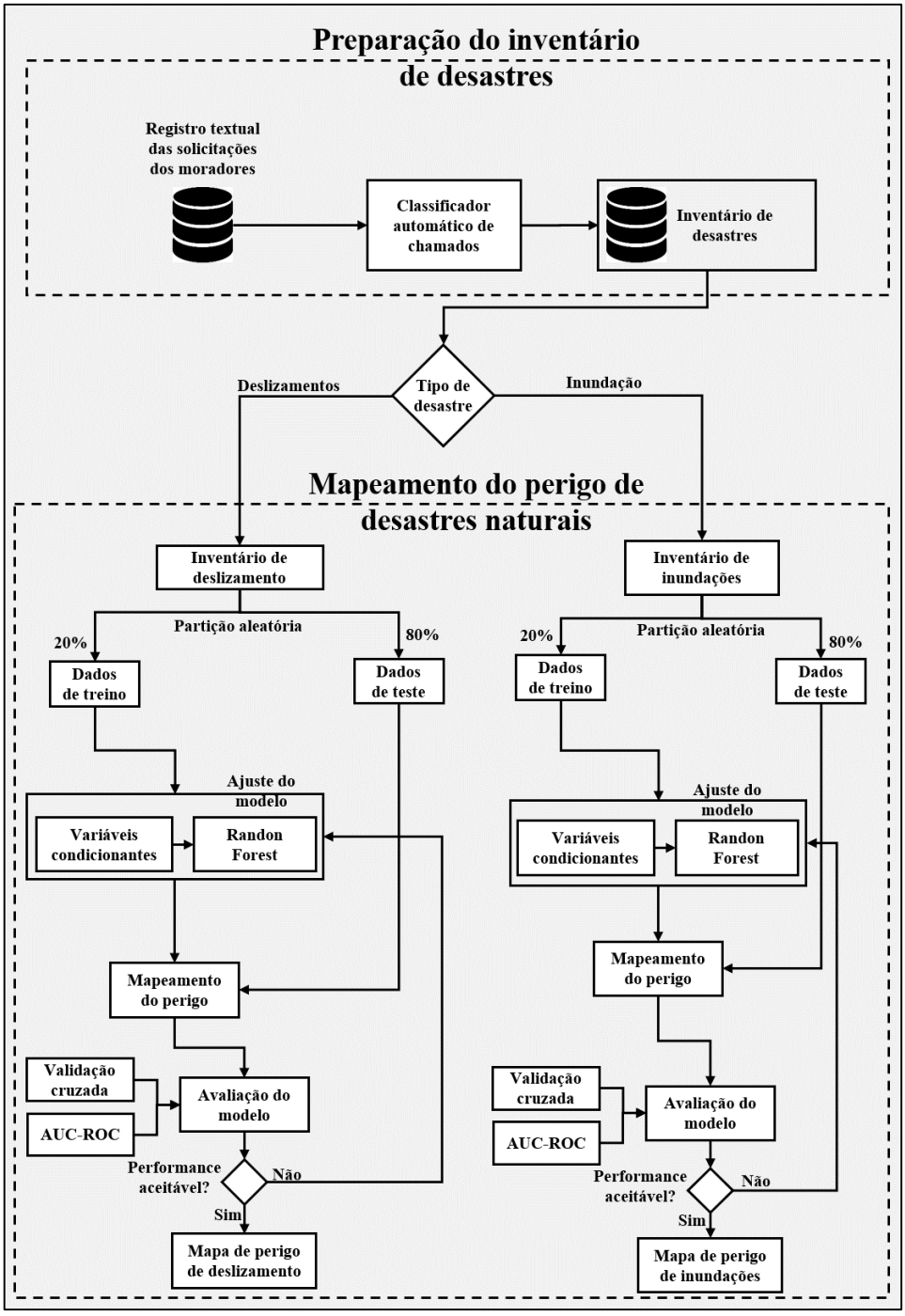
4.4 Conclusões do capítulo

No presente capítulo foi desenvolvida uma revisão sistemática da literatura sobre o tema de mapeamento de perigo de desastres naturais com algoritmos de aprendizado de máquina. Foi possível responder as questões de pesquisa estabelecidas com dados coletados em trabalhos científicos, fornecendo uma base sólida e confiável para comparação, com o objetivo de auxiliar novos estudos na área. Além disso, com o resultado final da revisão sistemática, foi estabelecido um processo padrão para o mapeamento de perigo de deslizamentos e inundações utilizando algoritmos de aprendizagem de máquina.

5 MODELO PROPOSTO

A presente seção tem como objetivo apresentar e discutir o modelo proposto para o mapeamento de perigo de desastres naturais. O mesmo foi dividido em duas fases interconectadas e interdependentes: criação do inventário de desastres e mapeamento do perigo de desastres naturais, conforme exposição do Fluxograma 8.

Fluxograma 8– Modelo proposto



Fonte: O Autor (2019)

Na fase de criação do inventário de desastres irá ocorrer a coleta dos dados textuais brutos, tratá-los, georreferenciá-los e classificá-los de acordo com o desastre específico. Antes de pôr o modelo proposto em uso é necessário realizar o ajuste do modelo e avaliação de performance. O algoritmo escolhido para realizar a classificação textual foi o *Naive Bayes*. Tal algoritmo foi escolhido por suas vantagens em termos de simplicidade, velocidade e eficiência, capacidade em lidar com sujeira nos dados e valores omissos, necessita de poucos dados para treinamento, e obter as predições é relativamente fácil (LANTZ, 2015). Além de todas essas vantagens tal algoritmo é largamente utilizado em tarefas de classificação de texto, e consegue obter acurácia igual ou superior a algoritmos mais sofisticados (BOUCKAERT, 2006; FRANK; PRANCKEVIČIUS; MARCINKEVIČIUS, 2017). Vale ressaltar que a etapa de ajuste só é realizada uma única vez e caso a performance seja adequada o modelo então é posto em produção.

Já na fase de mapeamento do perigo de desastres naturais o objetivo é utilizar os dados gerados na fase anterior para quantificar o grau de perigo de desastres naturais que cada localidade está exposta. Para isso, foi adotado o algoritmo *Random Forest* e variáveis condicionantes definidas com base na revisão sistemática da literatura, no capítulo 4 do presente trabalho. Após o ajuste dos modelos a performance será avaliada através de dois métodos comumente utilizados na literatura específica e caso esteja dentro dos padrões aceitáveis, o mapa de perigo será elaborado.

A escolha do algoritmo *Random Forest* foi feita através da análise dos resultados da revisão sistemática da literatura, bem como na comparação entre as vantagens e desvantagens de cada algoritmo expostas nos trabalhos analisados.

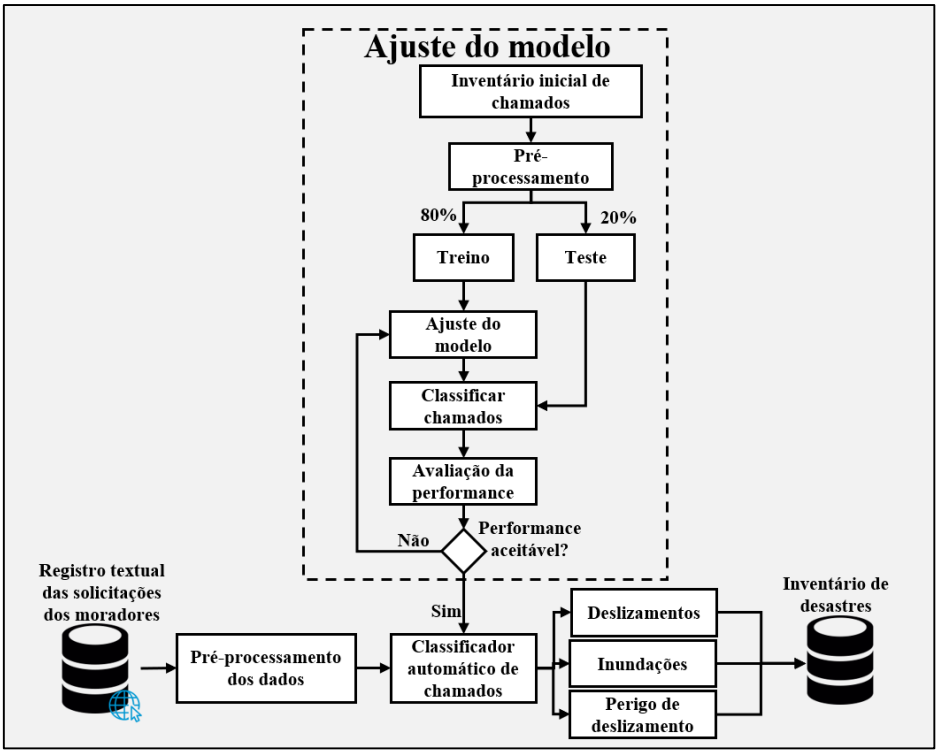
Random Forest é o algoritmo conjunto de classificação mais utilizado, aplicado em uma variedade enorme de problemas (AL-ABADI, 2018). Não são necessárias suposições a respeito dos dados e tanto variáveis numéricas como categóricas podem ser usadas, tornando-o adequado para modelar relações hierárquicas não lineares em grandes bancos de dados (LEE *et al*., 2017; SEGONI *et al*., 2015). Valores omissos nos fatores condicionantes podem ser tratados de forma flexível, dessa forma não é necessário remover casos com dados omissos (LEE *et al*., 2017). Wang e outros (2015c) argumentam que o Algoritmo Random Forest possui várias vantagens, incluindo alta acurácia, tolerância aceitável à *outliers* e ruídos nos dados, fácil eliminação de problemas como *over-fitting*, o qual é um problema comum em outros algoritmos, como por exemplo, ANN (AL-ABADI, 2018). Quando comparado com algoritmos como SVM, *Random Forest* tem a vantagem de fácil parametrização e melhor capacidade de generalização (FENG; LIU; GONG, 2015). Além disso, com base nos resultados da revisão

sistemática da literatura, das 25 vezes que o algoritmo *Random Forest* foi comparado com outros, obteve performance maior em 17, o que representa 68%. Tal resultado mostra que apesar de simples, possui alto poder preditivo e muitas vezes obtém performance melhor que algoritmos tradicionais. Um outro fator determinante para a escolha do algoritmo *Random Forest* foi a capacidade de calcular a importância de cada fator condicionantes presente no modelo (BEHNIA; BLAIS-STEVENS, 2018; ZHANG *et al*., 2017a).

5.1 Preparação do inventário de desastres

A fase de preparação do inventário representa o principal diferencial do presente estudo. Como visto na revisão sistemática da literatura, a principal fonte de dados para o mapeamento do perigo de desastres naturais são registros históricos dos eventos, bem como imagens provenientes de satélite. Entretanto, o presente estudo utiliza processamento de linguagem natural para extrair informações de registros textuais extraídos de chamados realizados por moradores à SEDEC e conseguinte criação do inventário de desastres naturais, contendo a descrição da solicitação, tipo do desastre e localização exata do evento. UM resumo é apresentado segundo o Fluxograma 9.

Fluxograma 9– Classificador automático de chamados



Fonte: O Autor (2019)

A atual fase pode ser dividida em duas etapas. A primeira é responsável pelo ajuste inicial do modelo de classificação de texto, utilizando o algoritmo *Naive Bayes*. Tal etapa é realizada uma única vez e a partir do momento que o algoritmo de classificação atingir a performance desejada não será mais necessário reajustar o modelo, salvo situações nas quais o objetivo seja implementar melhorias. A segunda etapa é quando o modelo ajustado é posto em produção. Pôr o modelo em produção significa que o mesmo será utilizado para realizar a classificação de novos chamados.

5.1.1 Ajuste do modelo de classificação de texto

A primeira tarefa da presente etapa é coletar os dados que serão utilizados no modelo de classificação de texto. Os dados utilizados no presente estudo, a título de exemplo, foram coletado através de uma consulta personalizada no portal de dados abertos da cidade de Recife-PE, disponibilizados pela (SEDEC, 2019). Vale ressaltar que a abordagem proposta pode ser empregada com qualquer dado semiestruturado em forma textual, desde que com o conteúdo do texto seja possível identificar algum tipo de desastre.

No caso do conjunto de dados utilizado nesse estudo, para ser gerado um registro é necessário que um morador entre em contato com a SEDEC, que então realizará o atendimento e irá registrar as informações em sistema. Todas as informações são registradas através de digitação das informações por um atendente, o que torna as informações vulneráveis a erros de digitação, ortografia, entre outros. No momento do atendimento são registradas informações sobre a descrição da solicitação, endereço, data da solicitação e se houve vítimas fatais.

Os chamados serão classificados em quatro classes mutuamente excludentes. A primeira classe representa os chamados que informam sobre eventos de deslizamentos já ocorridos ou em andamento. A classe inundação representa chamados relacionados a inundações já ocorridas ou em andamento. A terceira classe, perigo de deslizamento, representa os chamados que informam a respeito de situações potencialmente danosas, descreve eventos que possam vir a causar deslizamentos. Por fim, chamados diversos não relacionados a nenhuma das classes anteriormente citadas foram agrupados na classe “outros”. No Quadro 7 estão expostos exemplos de chamados na forma original para cada uma das classes utilizadas no presente estudo.

O pré-processamento envolve uma série de tarefas para tornar os dados compreensíveis e padronizados para o algoritmo. As tarefas preliminares realizadas foram a conversão para caracteres minúsculos, remoção de acentuação, números e qualquer caractere indesejado. Após

isso, foram aplicadas as etapas de pré-processamento que compreendem a eliminação de *stop words*, estemização, criação dos tokens e seleção dos termos mais importantes.

Quadro 7– Chamados classificados

Classe	Exemplo
Deslizamento	"34493812/A USUARIA PEDE A VISTORIA DA BARREIRA COM URGENCIA POIS ESTA DESLIZANDO, E O MURO ESTA RACHANDO."
Inundação	"AFIRMA QUE CERCA DE 200 RESIDENCIAS ENCONTRAM/SE DESOCUPADAS DEVIDO A ALAGAMENTO COM MAIS DE UM METRO DE AGUA DENTRO DAS RESIDENCIAS"
Perigo de deslizamento	"964USUÁRIA DESEJA AVALIAÇÃO DE BARREIRA E COLOCAÇÃO DE LONA COM BREVIDADE DEVIDO AO RISCO"
Outros	"3222/1135/O USUARIO PEDE A VISTORIA DO PREDIO,POIS A PAREDE ESTA ESQUENTANDO MUITO,O MESMO ACHA QUE É DEVIDO A CAIXA D´AGUA.PEDE URGENCIA."

Fonte: (SEDEC, 2019)

Uma outra tarefa importante na etapa de pré-processamento é a obtenção das latitudes e longitudes referentes a cada evento. Para isso, foi utilizada a Interface de Programação (*Application programming interface* – API) do Google disponível na linguagem R: Google *Geocoding*. Essa API toma como entrada o endereço e realiza uma busca no *Google Maps* e retorna a latitude, longitude e o endereço coletado, por extenso. Com isso foi possível obter as localizações exatas de cada evento, que servirá posteriormente para o mapeamento do perigo de desastres naturais.

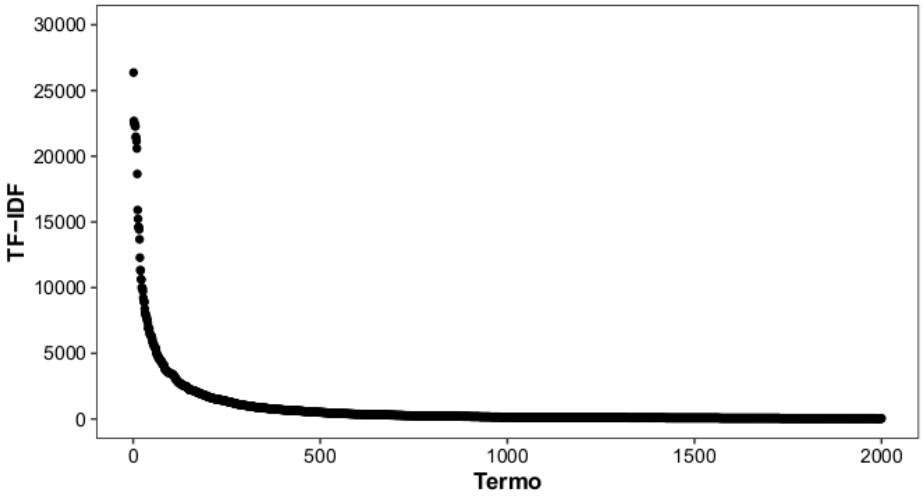
Vale ressaltar que devido a erros de digitação, erros na coleta dos endereços através da API eram esperados. Para filtrar os chamados cujos endereços foram coletados erroneamente utilizou-se uma filtragem na qual os endereços coletados fora dos limites geográficos foram excluídos. Tal método de filtragem tem como pressuposto que em uma cidade específica não existem duas ruas com nomes iguais.

O inventário inicial de chamados continha 40792 registros realizados no período entre 04/04/2012 e 08/05/2019. Após eliminar as observações com erros na coleta, o inventário ficou com 40211 chamados recebidos no período entre 04/04/2012 e 08/05/2019. Isso quer dizer que apenas 581 não foram localizados dentro dos limites da cidade em questão, o que representa 1,42% dos chamados. Vale destacar também que a abrangência temporal dos chamados não foi afetada.

Para selecionar os termos importantes foi calculado o *TF-IDF*, segundo a Equação (3.2). O Gráfico 14 mostra os termos em ordem decrescente do *TF-IDF*, para simplificação da visualização só foram exibidos os dois mil primeiros termos. É possível perceber que a partir

do termo 500 não há diferença significativa nos valores, dessa forma, para esse estudo serão selecionados apenas os 500 termos com maior *TF-IDF*.

Gráfico 14– TF-IDF para o conjunto de documentos



Fonte: O Autor (2019)

Os 20 termos com maior *TF-IDF* estão expostos no Tabela 6. É possível observar que os termos mais comuns são aqueles relacionados a alguma situação de perigo, como por exemplo os termos: barre, lon, desliz, entre outros. Tais termos representam as palavras barreiras, lon, deslizamento e quais outras palavras que possuam o mesmo tronco.

Tabela 6 – Os 20 termos com maior TF-IDF

Termo	TF-IDF	Termo	TF-IDF
barre	26368,25	**risc**	18651,93
usuar	22686,24	**desliz**	15904,8
lon	22479,32	**mesm**	15227,36
inform	22463,85	**mur**	14606,35
solicit	22281,93	**residenc**	14595,66
vist	22277,93	**plas**	14419,34
ped	21483,77	**local**	13673,6
usu	21373,92	**rachad**	12281,24
urgenc	21138,89	**monitor**	11329,92
colocac	20587,47	**verific**	11311,21

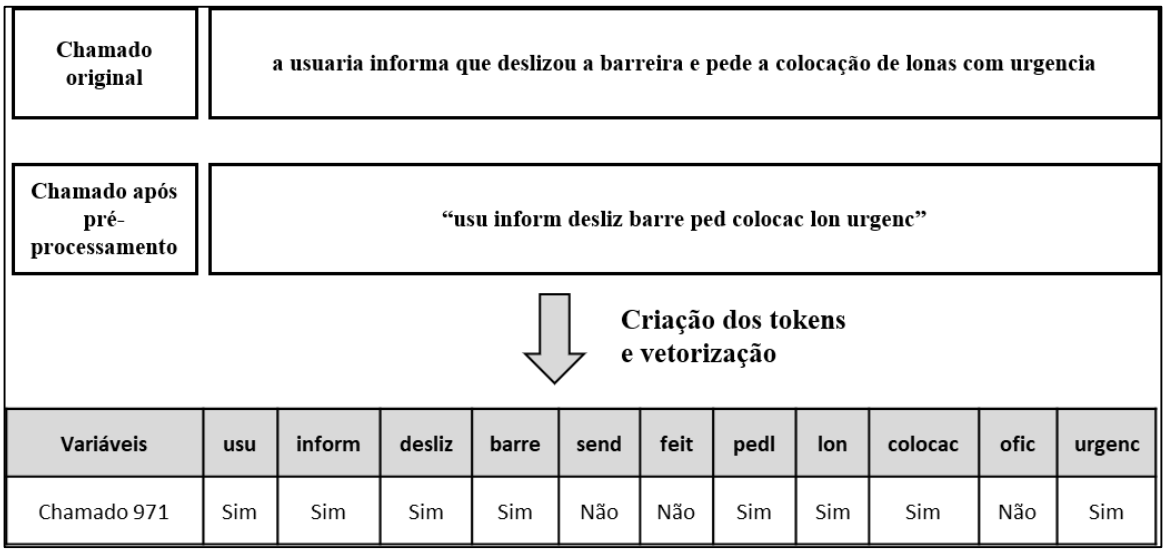
Fonte: O Autor (2019)

Uma vez realizado o pré-processamento foi possível criar os dados de treino e teste do modelo, na proporção 80% e 20%, respectivamente. No modelo ajustado no presente estudo,

entende-se por *tokens* as palavras restantes nos chamados após o pré-processamento. Tais *tokens* foram utilizados como variáveis. Dessa forma, cada observação foi transformada em um vetor de tamanho *M*, onde *M* representa o número de *tokens* identificados na amostra e cada elemento do vetor indica se o token em questão está ou não presente naquela observação. Veja o exemplo exposto na Figura 9.

Inicialmente é exibido o chamado na forma como foi coletado, logo em seguida é exibido como o chamado fica após a fase de pré-processamento e por fim, a forma vetorial que é utilizada no algoritmo *Naive Bayes*.

Figura 9– Exemplo do formato dos dados utilizados no algoritmo Naive Bayes



Variáveis	usu	inform	desliz	barre	send	feit	pedl	lon	colocac	ofic	urgenc
Chamado 971	Sim	Sim	Sim	Sim	Não	Não	Sim	Sim	Sim	Não	Sim

Fonte: O Autor (2019)

O algoritmo *Naive Bayes* foi ajustado e obteve bons resultados de performance, como exposto na Tabela 7. Como é possível observar, a taxa de falsos positivos é bem maior que a taxa de falsos negativos. Isso significa que o modelo tem uma probabilidade maior de classificar chamados que não são referentes a nenhum tipo de desastres como chamados referentes a desastres. Por outro lado, a probabilidade de o modelo classificar chamados referentes a desastres como não referente a nenhum tipo de desastre é muito baixa. Na prática, esse resultado tem grande importância, pois o custo de falsos negativos é muito grande, visto que não irá indicar desastre, quando na verdade há.

Um outro ponto importante que deve ser observado são as taxas de acerto em cada classe. Perceba que para as classes dos desastres de interesse, a taxa de acerto é maior que 90%, isso quer dizer o modelo foi capaz de classificar mais de 90% dos chamados corretamente. Em

segundo lugar, a taxa de acerto para a classe de risco de deslizamento obteve valor de 73,6%. Esse valor baixo ocorreu devido ao algoritmo não conseguir separar bem risco de deslizamento e deslizamento, classificando muitos chamados da classe de risco de deslizamento como deslizamentos.

Tabela 7 - Avaliação da performance do algoritmo Naive Bayes

Métrica	***Valor***
Acurácia	*0,8671*
Sensibilidade	*0,8232*
Especificidade	*0,9518*
Taxa de falsos positivos	*0,2644*
Taxa de falsos negativos	*0,0293*
Índice Kappa	*0,8038*
Índice de acerto de deslizamentos	*0,9090*
Índice de acerto de Inundação	*0,9380*
Índice de acerto de risco de deslizamento	*0,7360*

Fonte: O Autor (2019)

Uma vez que o modelo foi ajustado e a performance alcançou os patamares desejados, é possível utilizá-lo para realizar as classificações de chamados futuros. Nesse ponto, os chamados serão imputados no modelo e serão classificados em quatros classes discutidas anteriormente. Após a classificação, os chamados classificados em um dos tipos de desastres de interesse serão armazenados no banco de dados para compor o inventário de desastres.

5.1.2 Produção do modelo

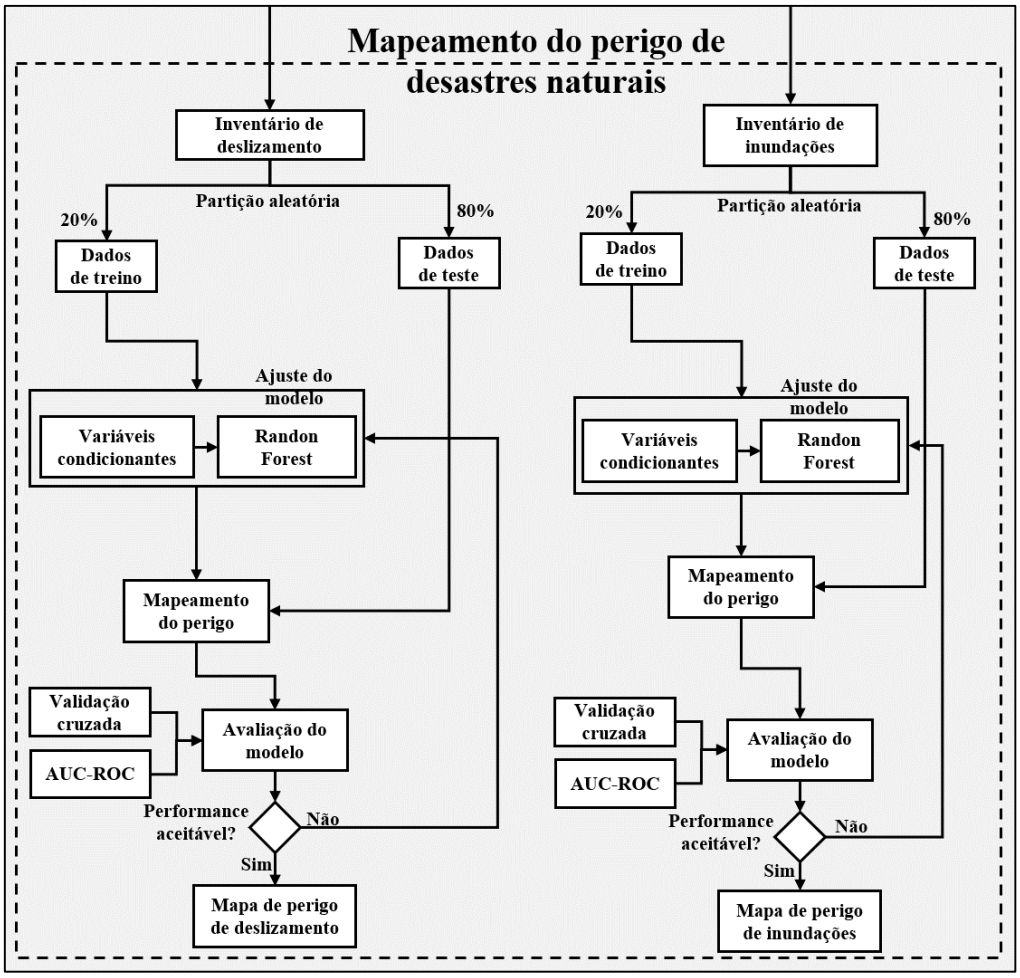
Uma vez que o modelo de classificação de texto foi ajustado e sua performance atingiu os patamares adequados, tal modelo pode ser integrado na fase de produção. Quando novos chamados forem adicionados e coletados do banco de dados disponibilizado por SEDEC (2019), esses passarão pelas mesmas etapas de pré-processamento realizadas no momento de ajuste do modelo e classificação. Após a etapa de pré-processamento, os chamados serão classificados dentro de uma das classes estabelecidas.

É importante que as tarefas do pré-processamento sejam respeitadas, de forma que a estrutura que será fornecida para o classificador seja idêntica a estrutura usada na etapa de treinamento do modelo, caso contrário o classificador não será capaz de realizar a classificação. Para garantir isso, foram criadas funções na linguagem R de modo que para cada nova classificação, as mesmas etapas de pré-processamento sejam realizadas. Tais funções foram exaustivamente testadas e comprovaram sua eficácia.

5.2 Mapeamento do perigo de desastres

A fase de mapeamento do perigo de desastres é responsável por calcular o perigo para cada um dos desastres estudados. No inventário de desastres, os eventos estão identificados segundo o tipo do desastre e contém as informações sobre a localização geográfica, o que permite a integração com um GIS. A forma dos registros são arquivos vetoriais no formato de ponto. Um resumo visual completo está exposto no Fluxograma 10.

Fluxograma 10– Fase de Mapeamento do perigo de desastres naturais



Fonte: O Autor (2019)

Dado que cada evento registrado no inventário está devidamente identificado, é possível ajustar dois modelos, um para cada tipo de desastre. Cada inventário é então particionado na proporção 80% para treino e 20% para teste, após essa tarefa, tem início a etapa de ajuste do modelo.

5.2.1 Ajuste do modelo

5.2.1.1 Variáveis condicionantes

Para descrever o problema completamente, é necessário um conjunto representativo de variáveis condicionantes, específicas para cada tipo de desastre. Tais variáveis foram identificadas através da revisão sistemática da literatura levando em consideração a especificidade do problema e a disponibilidade de dados. Todas as variáveis serão tratadas como camadas no ambiente GIS. Vale ressaltar que para cada tipo de desastres, conjuntos de variáveis diferentes serão utilizados. O Quadro 8 exibe as variáveis utilizadas e as informações referentes a escala estatística dos dados utilizados, o formato do arquivo, a fonte, escala de referência de cada variável e o tipo de desastre para o qual a variável será utilizada.

Quadro 8 – Informações sobre os parâmetros utilizadas

Parâmetros	Escala	Formato	Escala/precisão	Desastre	Fonte
Declividade	Razão	Raster	30m	Deslizamento e Inundação	(INPE, 2019)
Aspecto	Nominal	Raster	30m	Deslizamento	(INPE, 2019)
Litologia	Nominal	Vetorial	1:25000	Deslizamento e Inundação	(CPRM, 2019)
Altitude	Razão	Raster	30m	Deslizamento e Inundação	(INPE, 2019)
TWI	Razão	Raster	30m	Deslizamento e Inundação	Derivação
Uso do solo e cobertura do solo	Nominal	Raster	1:100.000	Deslizamento e Inundação	(MAPBIOMAS, 2019)
Distancia para as rodovias	Razão	Raster	30m	Deslizamento	Derivação
Curvatura vertical	Ordinal	Raster	30	Deslizamento e Inundação	(INPE, 2019)
Curvatura horizontal	Ordinal	Raster	30	Deslizamento e Inundação	(INPE, 2019)
Propensão hidrogeológica	Nominal	Vetorial	1:25000	Deslizamento e Inundação	(CPRM, 2019)
Número de escoamento	Intervalar	Vetorial	1:2 50.000	Deslizamento e Inundação	(ANA, 2019)
Distância para cursos d'água	Razão	Raster	1:250.000	Inundação	(ANA, 2019)
SPI	Razão	Raster	30	Inundação	Derivação
NDVI	Razão	Raster	30	Deslizamento e Inundação	Landsat 8

Fonte: O Autor (2019)

A declividade indica o ângulo em relação ao solo e pode ser expressa em ângulo, entre 0° e 90°, ou em porcentagem, que pode variar de 0 ao infinito. A declividade está representada no formato de porcentagem no presente trabalho. Tal variável é usada em vários estudos para o

mapeamento de perigo de desastre, foi utilizada nos trabalhos de Pham e outros (2018), Tien Bui e outros (2012a), Lee e outros (2017), Tehrany e outros (2015), entre outros.

O aspecto, que também é conhecido como orientação de vertentes, é definida como o ângulo azimutal correspondente a maior inclinação do terreno, no sentido descendente (INPE, 2019). Tal variável pode ser expressa numericamente em graus, entre 0° e 360°, bem de forma categórica em octantes, formato adotado no presente estudo. Tal variável foi utilizada nos estudos publicados por Ada E San (2018), Yao e outros (2008), Lee e outros (2018).

A formação rochosa do solo é representada pela variável litologia. Tal variável assume a forma categórica, pois para cada observação fornecerá o tipo da rocha que deu origem ao solo. Essa variável também é muito utilizada em mapeamento de perigo de desastres naturais, pois fornece relevantes informações sobre o solo. Tal variável foi utilizada nos trabalhos de Tien Bui e outros (2016a), Razavi Termeh e outros (2018), Wang e outros (2017).

A altitude expressa a altura do solo em relação ao nível do mar, em metros. Tal variável é uma das mais básicas das variáveis topográficas e é amplamente utilizada nos estudos de análise de perigo de desastres naturais, como por exemplo no trabalhos publicados por Solaimani e outros (2013), Pham e outros (2017).

O Índice Topográfico de Umidade (*Topographic wetness index* – TWI) é um modelo conceitual simplificado do processo hidrológico o qual prover indicação a respeito das características de umidade topográfica Nandi e outros (2016). O valor da variável pode ser calculado através da Equação (5.1), segundo Beven e Kirkby (1979). Onde α representa a acumulação de fluxo para uma determinada área e β representa a respectiva declividade. Tal variável foi utilizada em diversos estudos, como por exemplo: Zhang e outros (2017a), (Hong e outros (2015), shafizadeh-moghadam e outros (2018).

$$TWI = ln\left(\frac{\propto}{tan\,\beta}\right) \tag{5.1}$$

O uso do solo descreve os padrões de utilização do solo em determinada localidade. As informações providas por essa variável ajudam a entender a quais modificações o solo foi submetido. O tipo da variável utilizada é categórico. Da mesma forma das demais variáveis, o uso do solo foi usado em vários estudos para a caracterização do solo, como os trabalhos publicados por Nandi e outros (2016), Lee (2007), Giovannettone e outros (2018).

A distância para as rodovias tem como objetivo quantificar a distância entre os eventos e as rodovias, em metros. Tal variável foi selecionada devido as alterações provocadas na estrutura do solo devido as técnicas de construção de estradas. Tal variável é amplamente

utilizada e com importância reconhecida, como mostram os estudos publicados por Su e outros (2015), Feizizadeh e outros (2017), Chen e outros (2018c), entre outros.

A curvatura do terreno tem influência direta no processo de formação de deslizamento e inundações. Assim sendo, as variáveis selecionadas para representar a forma do terreno foram a curvatura vertical e curvatura horizontal. Curvatura vertical expressa o formato da vertente quando observada em perfil, pode ser entendido como a variação da declividade, formando padrões côncavo/convexo do terreno (VALERIANO, 2008). Tal variável assume o formato categórico, variando desde côncavo até convexo em 5 categorias, a saber: muito côncava, côncava, retilíneo, convexo e muito convexa. Por sua vez, a curvatura horizontal expressa o formato da vertente quando observada em perspectiva horizontal, na percepção comum pode ser traduzida no caráter de convergência ou divergência da linha de fluxo. De forma semelhante, a curva horizontal é do tipo categórica com 5 categorias, variando de muito convergente até muito divergente (VALERIANO, 2008). Juntas essas duas variáveis fornecem informações sobre a forma do terreno, que quando combinadas dão origem a 9 categorias. Estudo publicados por Chen e outros (2018a), Balamurugan e outros (2016), Hong e outros (2017a) utilizaram tais variáveis para caracterizarem a forma do terreno.

A propensão hidrogeológica informa a quais eventos determinada região está sujeita. Tal variável foi coletada através de estudos prévios realizados pela Agência Nacional de Águas (ANA).

O número de escoamento, também conhecido como *Curve Number* (CN) é um parâmetro empírico utilizado para obter previsões do volume de escoamento superficial. Quanto maior for seu valor, maior o volume de escoamento superficial formado. Tal método considera que a lâmina de escoamento superficial é função da precipitação e de perdas devido a infiltração no solo, obstáculos vegetais e bloqueios por terrenos (TYAGI *et al*., 2008). Os valores do número de escoamento podem ser obtidos através de tabelas padrões, métodos gráficos e ainda estimações (OLIVEIRA *et al*., 2016). Os valores utilizados no presente estudo foram disponibilizados pela ANA (ANA, 2019).

A distância para os cursos d'água é a medida da menor distância entre um evento i, e o curso d'água mais próximo, em metros. Foram considerados como cursos d'água rios e córregos. Tais cursos d'água foram identificados através de arquivos vetoriais disponibilizados pela (ANA, 2019). Estudos publicados por Rahmati e PourghasemI (2017), Al-Abadi (2018), Gaidzik e outros (2017), Wang e outros (2013) utilizaram tal variável para o mapeamento de perigo de inundações e deslizamentos.

Índice de Poder Erosivo (*stream power index* – SPI) é usado para mensurar o poder erosivo do fluxo de água, o qual é inversamente proporcional ao valor de estabilidade da declividade Althuwaynee e outros (2014). O SPI pode ser usado para descrever o potencial erosivo da correnteza em um dado ponto da superfície Park e Lee (2014). A medida que a área de captação e a declividade aumentam, a quantidade e velocidade do fluxo de água aumentam e, consequentemente o SPI também aumenta (FLORINSKY, 2012). Moore e outros (1991) definem o SPI segundo a Equação (5.2), onde A_s representa a área de captação de um determinado ponto e β o ângulo de declividade.

$$SPI = A_s \tan \beta \tag{5.2}$$

A cobertura do solo é uma variável categórica que informa a respeito da cobertura de um determinado ponto, seja cobertura natural ou artificial (asfalto, por exemplo). Vários autores utilizam tal variável para modelar o comportamento de um determinado ponto em relação a eventos naturais, como por exemplo os trabalhos publicados por Sciarra e outros (2017), (Peng e outros (2014), Ermini e outros (2005).

Mapas obtidos a partir de imagens de satélite mostram a densidade do crescimento das plantas em todo o globo terrestre, tendo como principal o Índice Normalizado de Diferença da Vegetação (*Normalized Difference Vegetation Index* – NDVI) (YILMAZ, 2010a). Segundo Xia e outros (2017), o NDVI pode ser calculado através das bandas disponíveis nas imagens Landsat 8 UNITED STATES GEOLOGICAL SURVEY - USGS (2019), segundo a Equação (5.3). Yilmaz (2010a) explica que valores muito pequenos do NDVI (abaixo de 0,1) correspondem a áreas áridas, areia ou neve. Valores moderados (0,2 – 0,3) correspondem a pequenos arbustos ou pastagem. Por fim, valores elevados (0,6 – 0,8) correspondem a florestas tropicais.

$$NDVI = \frac{(banda\ 5 - banda\ 4)}{(banda\ 5 + banda\ 4)} \tag{5.3}$$

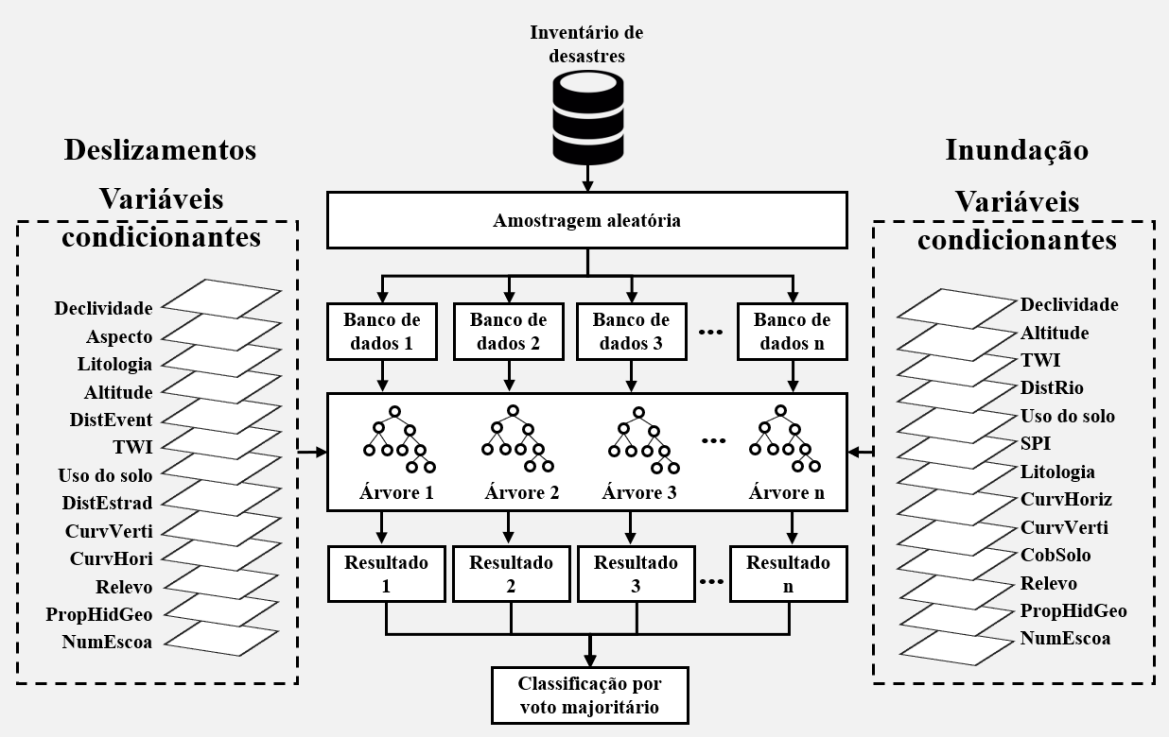
Todas as variáveis citadas anteriormente foram tratadas no ambiente de um software GIS. Para cada ponto correspondente a um desastre, foram coletados os valores equivalentes para cada uma das variáveis e então usadas para o treino do algoritmo *Random Forest*.

5.2.1.2 Ajuste do algoritmo Random Forest

O algoritmo *Random Forest* (BREIMAN, 2001) foi implementado na linguagem R. Para cada tipo de desastre foi ajustado um modelo. A tarefa do modelo é classificar, um dado ponto no espaço, em duas classes: classe positiva (1) ou classe negativa (0). Ser classificado na classe

positiva significa que aquele ponto em específico está sujeito a um desastre. Por sua vez, ser classificado na classe negativa significa que o determinado ponto não está sujeito ao desastre. Na Figura 10 está exposto o modelo utilizado nesse trabalho.

Figura 10– Modelo de mapeamento de perigo de desastres naturais



Fonte: O Autor (2019)

Como explicado anteriormente, o algoritmo *Random Forest* é formado com uma combinação de *n* árvores. Dessa forma cada ponto submetido ao modelo será classificado como pertencendo a classe positiva ou negativa, em cada uma das *n* árvores. Após a classificação por todas as árvores, o perigo será igual à proporção de árvores que o classificaram como pertencente a classe negativa. Por exemplo, imagine que um determinado ponto foi submetido a um modelo *Random Forest* com cem árvores, setenta o classificaram na classe positiva e o restante na classe negativa. Consequentemente o perigo de desastre nesse ponto é de 0,7.

Tal abordagem é comum para o mapeamento de perigo de desastres naturais. (Hong e outros (2016) compararam o método com outras técnicas e concluíram que a abordagem que utiliza o *Random Forest* consegue atingir acurácia suficiente. Por sua vez, Wang e outros (2015b) mapearam o perigo de inundações utilizando o método citado. Da mesma forma Feng

e outros (2015), mapearam o perigo de inundações, atingindo acurácia de 0,94 e índice Kappa 0,88.

5.2.2 Avaliação da performance

Após o algoritmo ser treinado, a performance do modelo será avaliada utilizando os dados de teste. Em primeiro lugar é necessário realizar as classificações para os dados de teste, para então a performance ser avaliada. Os métodos utilizados para a avaliação da performance serão a validação cruzada e a AUC. A performance obtida será comparada com o padrão estabelecido na revisão sistemática da literatura e caso seja suficientemente alto o modelo poderá então ser utilizado para a produção de mapas de perigo.

5.2.2 Produção do mapa de perigo

Para a produção do mapa de perigo será criado um arquivo vetorial de pontos com diferença de 30 m entre eles. Após a geração dos pontos regulares, as mesmas variáveis utilizadas para o treino do modelo serão coletadas para a amostra de pontos gerados. Cada ponto será então classificado pelo modelo e o perigo calculado para cada tipo de desastre. Após o cálculo do perigo, os pontos serão transformados em arquivos raster (tipo de arquivo matricial utilizado no GIS) e dado o tratamento necessário para a melhor visualização. A coleta dos valores das variáveis, tratamento, geração do arquivo *raster* e confecção dos mapas serão realizados no Qgis (QGIS DEVELOPMENT TEAM, 2019). Já a classificação dos pontos será realizada na linguagem R (R CORE TEAM, 2019).

5.3 Conclusões do capítulo

No presente capítulo os modelos propostos foram apresentados e discutidos. A estrutura do modelo de classificação textual foi apresentada e a performance avaliada com base em métricas de validação cruzada. A acurácia do modelo foi de 0,867, demonstrado bom valor preditivo. Além disso, as estruturas dos modelos de classificação de perigo foram apresentadas, bem como as variáveis condicionantes utilizadas. Por fim, os processos para geração dos mapas de perigo foram exibidos e discutidos.

6 ESTUDO DE CASO

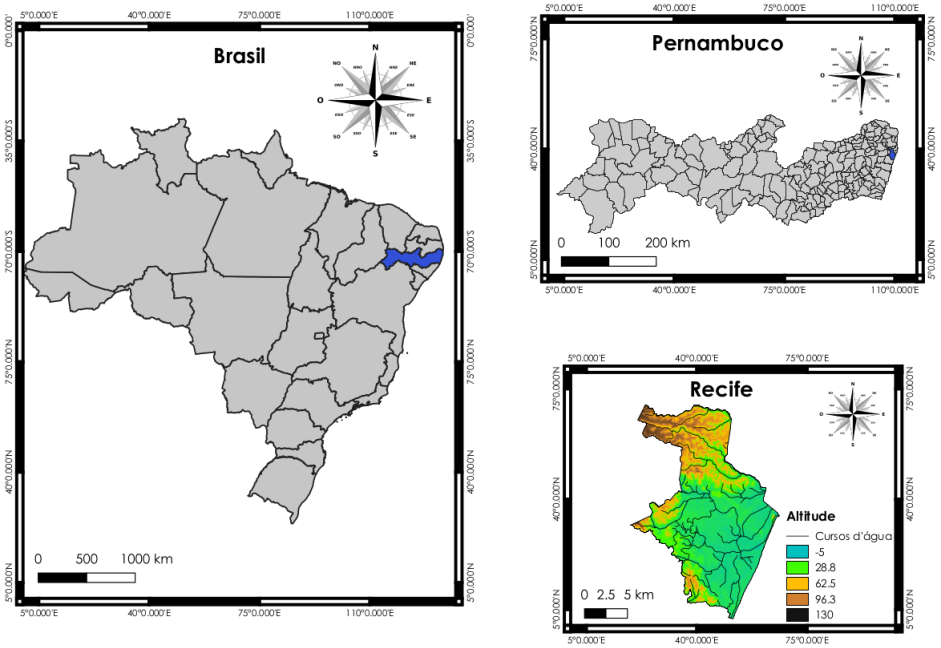
O atual capítulo tem como objetivo mostrar em detalhes a aplicação do modelo proposto para uma cidade situada na região nordeste do Brasil, no estado de Pernambuco. Inicialmente será realizada a delimitação da área de estudo e apresentação do inventário de eventos. Em seguida, serão apresentadas as variáveis condicionantes para só então, o modelo ser ajustado e seus resultados avaliados.

Como discutido na seção 4.3 do presente trabalho, uma importante tarefa quando os dados utilizados para treino do modelo são provenientes de fontes alternativas é a verificação dos resultados do modelo. Em outras palavras, os resultados serão comparados com registros de desastres históricos e relatório oficiais de órgãos competentes para verificar a capacidade do modelo de descrever a realidade dos eventos.

6.1 Delimitação da área de estudo

A cidade de Recife é capital do estado de Pernambuco, nordeste do Brasil. O terreno da cidade possui grandes variações de altitude e declividade, além de possuir uma grande malha de cursos d'águas, conforme exposto no Mapa 1. Recife tem população estimada para 2019 de mais de 1,5 milhões de pessoas com densidade demográfica de 7039 hab/km² (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE, 2011).

Mapa 1– Delimitação da área de estudo



Fonte: O Autor (2019)

O território da cidade ocupa um total de 218 km². O relevo é composto por 9 tipos diferentes, dos quais predominam os tabuleiros dissecados, ocupando aproximadamente 75 km² do território (CPRM, 2019). Na Tabela 8 está exposto todas as 9 formações, bem como as respectivas áreas. Já a formação litológica do solo é composta por 12 tipos de rocha, na qual a maior parte do território é formado por Arenito Conglomerático e Argilito Arenoso, totalizando 78,21 km², conforme exposição da Tabela 9 (CPRM, 2019). A declividade do terreno é formada em sua maioria por topos planos restritos, com ângulo variando entre 0 e 3° e vertentes variando de 10° a 25°, ocupando quase 75 km², como exposto na Tabela 10 (INPE, 2019).

Tabela 8 – Relevo da cidade de Recife

Relevo	Área (km²)
Colinas	10,73
Corpo d'água	5,90
Morros Baixos	2,21
Planícies Fluviais (planícies de inundação, baixadas inundáveis e abaciamentos)	10,29
Planícies Fluviomarinhas (brejos)	41,52
Planícies Flúviomarinhas (mangues)	6,47
Planícies Marinhas (terraços marinhos e cordões arenosos)	60,54
Tabuleiros	5,63
Tabuleiros Dissecados	74,96

Fonte: (CPRM, 2019)

Tabela 9 – Litologia da cidade de Recife

Litologia	Área (km²)
Areia	4,41
Areia, silte, argila	10,29
Arenito	2,84
Arenito arcoseano, Arenito conglomerático, Ritmito	1,59
Arenito conglomerático, Argilito arenoso	78,21
Arenito conglomerático, Conglomerado	0,03
Argila, matéria orgânica	6,47
Argila, silte, areia, bioclastos	56,13
Argila, silte,areia	41,52
Corpo d'água	5,90
Metadiorito, Migmatito, Ortognaisse granodiorítico, Ortognaisse granítico, Ortognaisse tonalítico	2,43
Ortognaisse	8,43

Fonte: (CPRM, 2019)

Tabela 10 – Declividade do terreno

Declividade	area
0 - 3º	10,29402
0 - 5°	60,53646
3 -10º	10,72955
5 - 20º	2,21418
Plano 0°	53,89506
Topo Plano: 0 - 3° Vertente: 10 - 25°	5,63219
Topos planos restritos: 0 - 3° Vertente: 10 - 25°	74,95904

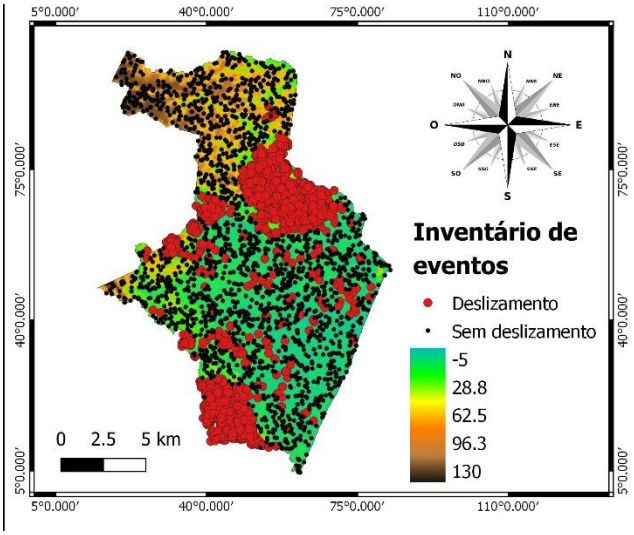
Fonte: (CPRM, 2019)

Devido às características morfológicas e geológicas do terreno, aliado com as chuvas, regimes das marés e a urbanização não planejada, a cidade de Recife sofre com diversos eventos de deslizamentos e inundação, como discutido em capítulos anteriores. Dessa forma, é uma excelente candidata para o estudo de caso que será desenvolvido nas seções subsequentes.

6.2 Inventário de eventos

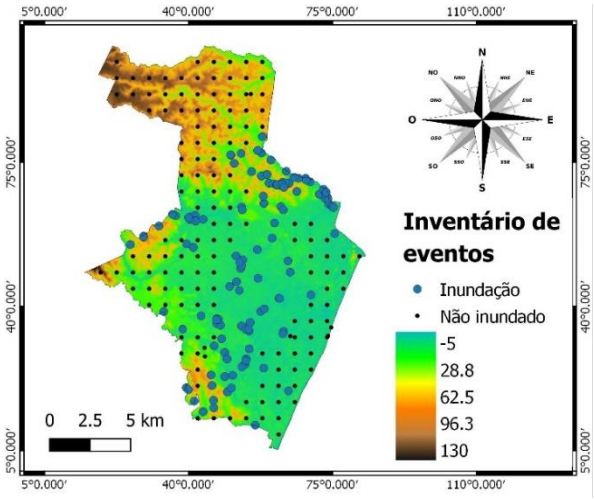
Com a aplicação do modelo descrito no capítulo 5 para classificação dos chamados em suas respectivas classes, foi possível criar um inventário de eventos para cada um dos tipos de desastres incluídos no presente estudo, conforme exposição do Mapa 2 e Mapa 3. Foram identificados 3468 pontos (centroides) de deslizamentos e 160 pontos (centroides) de inundação.

Mapa 2– Inventário de eventos de deslizamentos



Fonte: O Autor (2019)

Mapa 3– Inventário de eventos de inundações



Fonte: O Autor (2019)

A metodologia para gerar os pontos nos quais não houve desastres foi baseada na criação de polígonos. Para tal, os polígonos foram criados de maneira a cobrir todos os pontos nos quais houve desastre. Pontos aleatórios foram gerados fora dos polígonos para evitar a sobreposição de pontos onde não houve desastres com pontos onde houve desastres. A única diferença entre a metodologia usada para deslizamentos e inundações foi que os pontos para o treino de inundações foram gerados de forma regular, devido a baixa quantidade de pontos pertencentes a classe positiva.

Ao final foram gerados dois inventários de eventos. O primeiro referente aos eventos de deslizamentos, contendo 5386 pontos, sendo 64,39% dos pontos com deslizamentos. O segundo inventário, referente a inundações, contendo 308 pontos, com proporção de pontos de inundação de 51,94%.

É possível perceber que existe uma concentração alta de eventos de deslizamentos nas áreas com maior altitude, enquanto as inundações seguem um padrão que varia segundo a proximidade dos cursos d'água. Apesar de existir certa relação, a análise visual dos dados não é capaz de explicar todos os eventos, por isso o modelo de Aprendizado de Máquina será ajustado com o objetivo de capturar todas as interações não percebidas pelas técnicas convencionais.

6.3 Variáveis condicionantes

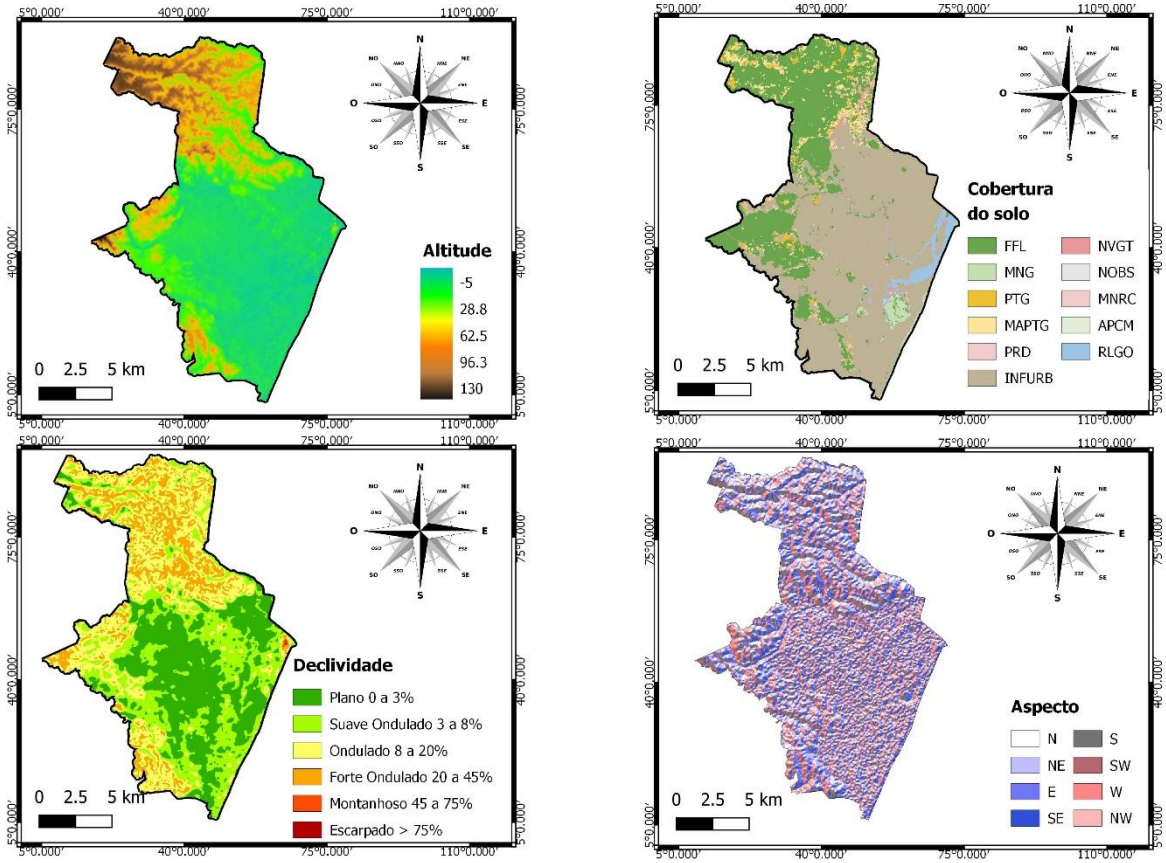
As variáveis condicionantes selecionadas e discutidas no capítulo 4 foram coletadas e tratadas para a área de estudo. Para isso, os dados brutos coletados, segundo exposição do

Quadro 8, foram tratados no ambiente GIS. Os tratamentos aplicados foram recorte para a área de estudo, projeção para o sistema de referência geográfica SIRGAS 2000, UTM ZONE 25S, e coleta dos valores para os pontos presente nos inventários de eventos.

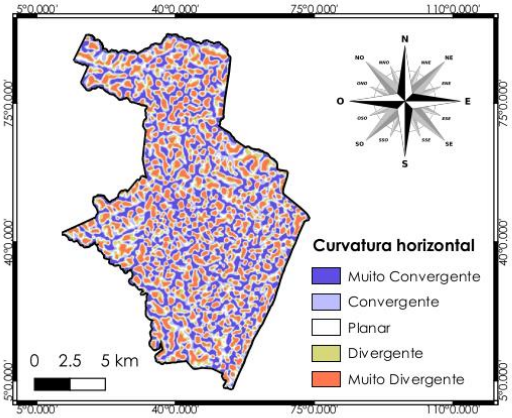
Ao todo foram utilizadas 14 variáveis condicionantes, conforme expostas no Mapa 4. Como exposto no Quadro 8, das 14 variáveis utilizadas no estudo, 12 delas foram utilizadas no modelo de deslizamento e 12 utilizadas no modelo de inundação.

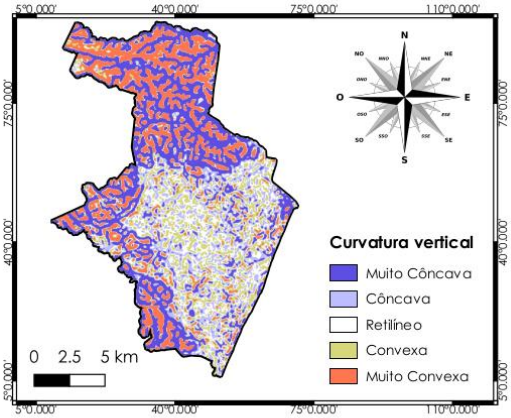
Os inventários foram repartidos em dois bancos de dados, um de treino e outro de teste. Para os dados de treino foram utilizados 80% dos dados originais, e o restante (20%) foi usado para teste do modelo. Vale ressaltar que para cada tipo de desastre foram criados banco de dados diferentes.

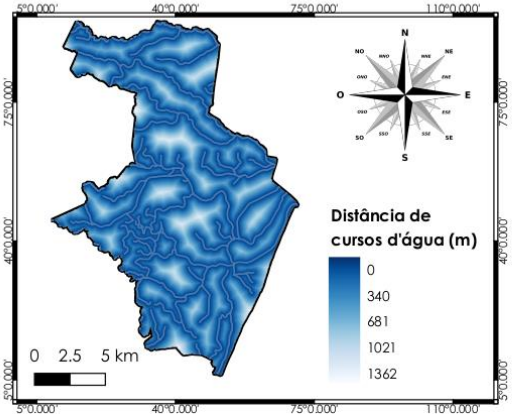
Mapa 4– Variáveis condicionantes a) altitude; b) cobertura do solo; c) declividade; e) relevo; f) curvatura horizontal; g) curvatura vertical; h) distância para cursos d'água; i) litologia; j) número de escoamento; l) propensão hidrogeológica; m) NDVI; n) SPI; o) TWI; p) distância para as rodovias

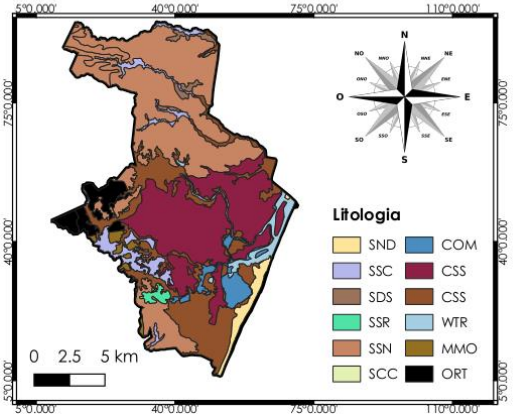


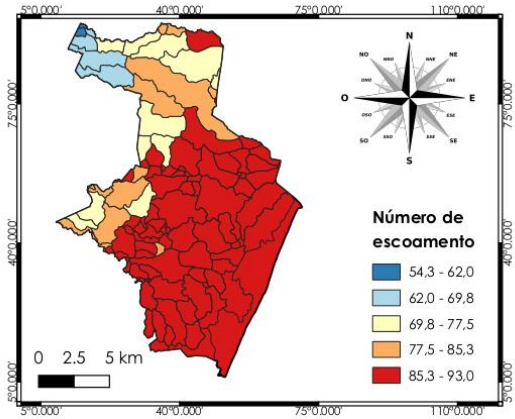
Mapa 4 - Continuação

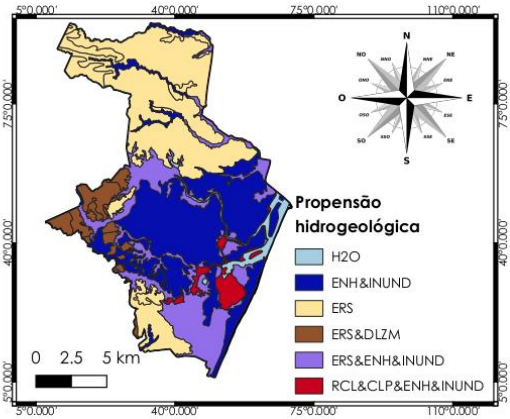

Curvatura horizontal
Muito Convergente
Convergente
Planar
Divergente
Muito Divergente
0 2.5 5 km

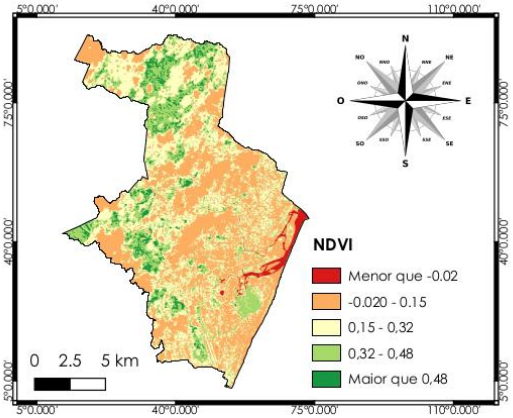

Curvatura vertical
Muito Côncava
Côncava
Retilíneo
Convexa
Muito Convexa
0 2.5 5 km

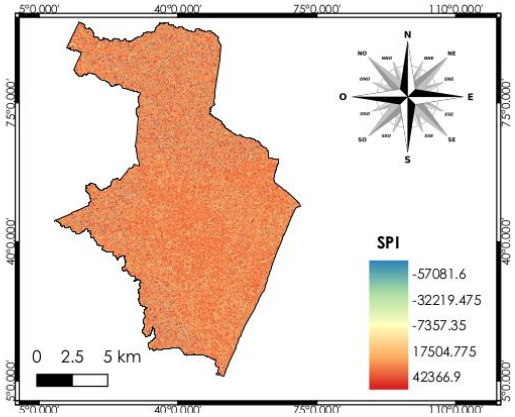

Distância de cursos d'água (m)
0
340
681
1021
1362
0 2.5 5 km


Litologia
SND
SSC
SDS
SSR
SSN
SCC
COM
CSS
CSS
WTR
MMO
ORT
0 2.5 5 km

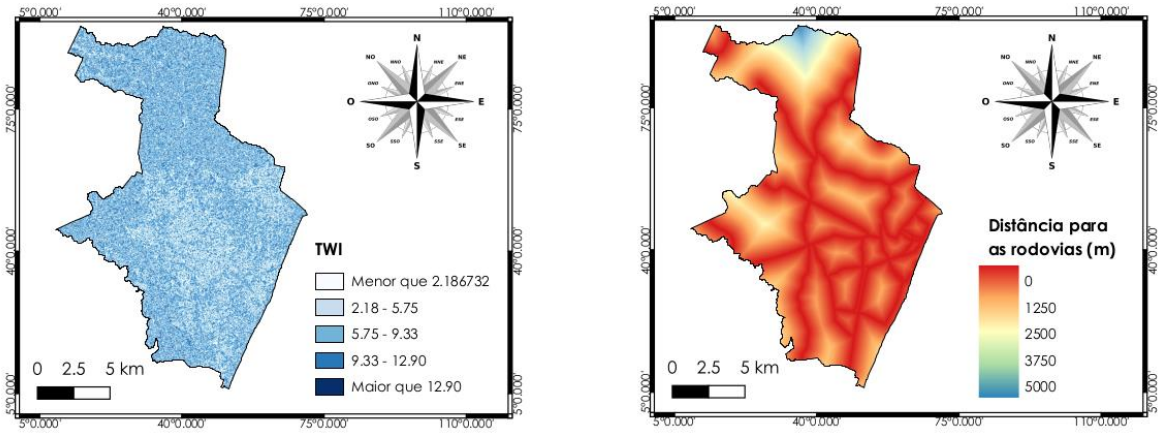

Número de escoamento
54,3 - 62,0
62,0 - 69,8
69,8 - 77,5
77,5 - 85,3
85,3 - 93,0
0 2.5 5 km


Propensão hidrogeológica
H2O
ENH&INUND
ERS
ERS&DLZM
ERS&ENH&INUND
RCL&CLP&ENH&INUND
0 2.5 5 km


NDVI
Menor que -0.02
-0.020 - 0.15
0,15 - 0,32
0,32 - 0,48
Maior que 0,48
0 2.5 5 km


SPI
-57081.6
-32219.475
-7357.35
17504.775
42366.9
0 2.5 5 km

Mapa 4 - Continuação



Fonte: O Autor (2019)

Como discutido previamente, as características morfológicas e hidrológicas da cidade de Recife são bastante diversificadas. Na Tabela 11 estão resumidas as principais características das variáveis condicionantes em relação aos pontos identificados como deslizamento e inundações.

Tabela 11 – Características das variáveis condicionantes (continua)

Parâmetros	Valores	Número de deslizamentos	Número de inundações
Declividade	Ondulado 8 a 20%	1680	50
	Forte Ondulado 20 a 45%	1079	12
	Plano 0 a 3%	491	46
	Suave Ondulado 3 a 8%	218	52
Aspecto	SW	521	18
	E	459	15
	S	457	11
	N	432	0
	NW	430	25
	SE	427	18
	W	387	16
	NE	355	23
Litologia	Arenito conglomerático, Argilito arenoso	2803	42
	Argila, silte, areia	300	75
	Argila, silte, areia, bioclastos	144	28
	Ortognaisse	102	7
	Areia, silte, argila	62	6
	Metadiorito, Migmatito, Ortognaisse granodiorítico, Ortognaisse granítico, Ortognaisse tonalítico	38	2
	Arenito arcoseano, Arenito conglomerático, Ritmito	14	0
	Areia	5	0

Tabela 12 – Continuação

Parâmetros	Valores	Número de deslizamentos	Número de inundações
Altitude	Entre 22 e 49m	1526	34
	Entre 49 e 76m	1194	15
	Menos que 22m	631	110
	Entre 76 e 103m	117	
	Entre 76 e 103m	0	1
TWI	Entre 10 e 13	1735	60
	Entre 2 e 6	1310	84
	Entre 6 e 10	214	7
	Maior que 13	133	8
	Menor que 2	76	1
Uso do solo e cobertura do solo	INFURB	3266	154
	MAPTG	119	6
	NVGT	56	0
	FFL	27	0
Distancia para as rodovias	Menor que 1 km	2129	-
	Entre 1 e 2 km	1338	-
	Entre 2 e 3 km	1	-
Curvatura vertical	MCCV	1830	76
	MCVX	1071	12
	CCV	233	32
	CVX	182	17
	RTL	152	23
Curvatura horizontal	MCVG	738	52
	CVG	704	44
	PLN	703	21
	MDVG	681	24
	DVG	642	19
Propensão hidrogeológica	Erosões	2785	39
	Erosões, Enchente e inundação	300	75
	Enchente e inundação	210	34
	Erosões, Deslizamentos	173	12
Número de escoamento	Entre 80 e 90	1623	57
	Maior que 90	1545	91
	Entre 70 e 80	300	12
Distância para cursos d'água	Menor que 270	-	115
	Entre 270 e 545	-	34
	Entre 545 e 820	-	8
	Entre 820 e 1100	-	3
SPI	Entre -17300 e 2600	-	106
	Entre 2600 e 22500	-	36
	Maior que 22500	-	12
	Entre -3700 e -17300	-	3
	Menor que -38000	-	3
NDVI	Entre 0 e 0.15	2125	112
	Entre 0,15 e 0,32	1271	47
	Entre 0,32 e 0,48	70	1

Fonte: O Autor (2019)

6.4 Treino e avaliação da performance do modelo

Os dados necessários para ajuste do modelo foram tratados utilizando a linguagem de programação R (R CORE TEAM, 2019). A primeira etapa consistiu no tratamento e adequação

dos dados para o treino do modelo, envolvendo atividades de limpeza e manipulação de variáveis. Nessa etapa foi utilizado o pacote *Tidyverse* (WICKHAM, 2017), que possui ampla variedade de ferramentas para manipulação, tratamento, visualização e modelagem, além de ser amplamente utilizado por diversos desenvolvedores. A segunda etapa, responsável pelo treinamento do modelo, foi utilizado o pacote *RandomForest* (LIAW; WIENER, 2002), uma implementação na linguagem R do código originalmente desenvolvido por Breiman (2001). Por fim, na etapa de avaliação da performance do modelo, foi utilizado o pacote *pROC* (ROBIN *et al*., 2011) para elaborar a curva AUC-ROC para avaliar a performance dos modelos.

Os modelos ajustados abordam a problemática de classificação. Ou seja, dado um ponto $Y_{ij} = \{X_{11}, X_{12}, X_{13}, \dots , X_{ij}\}$ onde X_{ij} representa o valor da variável j para no ponto i, a tarefa do modelo consiste em classificar Y_{ij} em duas classes mutuamente excludentes $C_i = \{0, 1\}$, de modo que 0 significa que o ponto i não está sujeito a desastre e 1 significa que o ponto i está sujeito ao desastre. Ao final, a proporção das árvores do modelo que classificaram o ponto i na classe 1 é calculada de modo que $0 \leq P_i \leq 1$ representa o perigo daquele ponto.

Os modelos foram ajustados com os mesmos parâmetros. Foram utilizadas 1000 árvores, com a escolha aleatória de 3 variáveis a cada corte. O erro *out-of-bag* para deslizamento foi de 4,21%, já para o modelo de inundação foi de 18,11%.

O erro *out-of-bag* dentro de cada classe para o modelo de deslizamento foi de 4% para ambas as classes, isso significa que o modelo de deslizamentos é capaz de classificar tanto na classe positiva quanto na classe negativa com a mesma precisão. Já para o modelo de inundação o erro *out-of-bag* para as classes variou. Para a classe positiva (classe 1) o erro foi de 10,40% e para a classe negativa (classe 0) foi de 26,27%.

O resultado obtido para o erro dentro da classe traz à tona uma importante discussão acerca do custo de cada tipo de erro. O modelo de inundação irá classificar em média 26,27% dos pontos que não estão sujeitos a desastres como pontos que estão sujeitos a desastres. Apesar do erro relativamente alto, o modelo foi considerado adequado pois o custo de se classificar eventos negativos na classe positiva é muito menor que o de classificar eventos positivos na classe negativa.

O custo de classificar um ponto positivo como negativo é que um ponto com verdadeiro risco de inundação seria negligenciado devido ao resultado errôneo, podendo levar a danos financeiros e a integridade humana. Por outro lado, a classificação errada de pontos negativos como positivos pode levar a tomada de decisões baseadas em informações erradas, levando a desperdício de recursos e ainda mais grave, deixar de atender um ponto realmente com alto risco.

A avaliação dos performance do modelo segundo as métricas derivadas da validação cruzada, expostas no Quadro 3, foram calculadas a partir dos dados de teste e apresentadas na Tabela 13. O modelo de deslizamento apresentou excelente desempenho, com acurácia acima da média e índice *Kappa* classificado como quase perfeito, segundo os critérios apresentados na Tabela 1. Já o modelo de inundação apresentou boa acurácia, bom poder de predição de positivos e negativos, porém, o índice *Kappa* o classifica como um modelo com poder preditivo moderado, segundo a Tabela 1.

Percebe-se que para o modelo de inundação, o poder de detecção dos negativos é maior que o poder de detecção dos positivos. Isso significa que o modelo ajustado tem maior erro ao detectar os pontos inundados. Um dos motivos para isso ter acontecido pode ter sido a pequena quantidade de dados positivos utilizados para treino do modelo. É esperado que com a coleta de mais pontos positivos o poder preditivo do modelo aumente.

O resultado da validação cruzada não invalida os resultados obtidos através do erro *out-of-bag*, pois variações dessa natureza são esperadas devido às características aleatórias do modelo *Random Forest*. Outro ponto importante a se notar é a quantidade de dados usados para teste do modelo de inundação foi baixa, devido a indisponibilidade, outro fator que influenciou para tal diferença. Os resultados dos erros *out-of-bag* foram calculados com uma quantidade maior de dados, por isso, sua confiança estatística é maior.

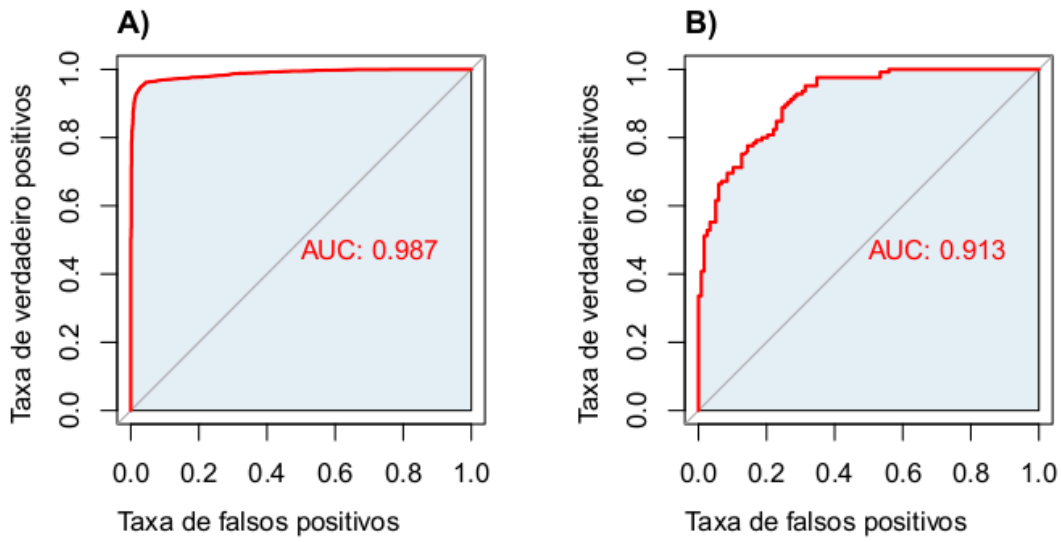
Tabela 13 – Avaliação da performance dos modelos segundo métrica de validação cruzada

Métrica	**Resultado**	
	Modelo de deslizamento	**Modelo de inundação**
Acurácia	0,9503	0,8
Sensibilidade	0,9421	0,8857
Especificidade	0,9646	0,70
Poder de predição de positivos	0,9789	0,775
Poder de predição de negativos	0,9049	0,84
Índice Kappa	0,8941	0,59

Fonte: O Autor (2019)

Um outro método utilizado para medir a performance dos modelos ajustados foi a AUC-ROC, que através da curva ROC mede a acurácia geral do modelo através do cálculo da AUC-ROC. As curvas ROC para os dois modelos ajustado estão expostas no Gráfico 15.

Gráfico 15– Curva ROC para o modelo de a) deslizamento b) inundação



Fonte: O Autor (2019)

Segundo a métrica de avaliação AUC-ROC, ambos os modelos obtiveram desempenho suficiente para serem colocados em produção. A AUC para o modelo de deslizamento foi de 98,70% e para o modelo de inundação de 91,30%.

Segundo o padrão estabelecido apresentado na Tabela 2, os modelos ajustados no presente estudo estão com desempenho adequado, quando comparado com os modelos publicados em periódicos científicos.

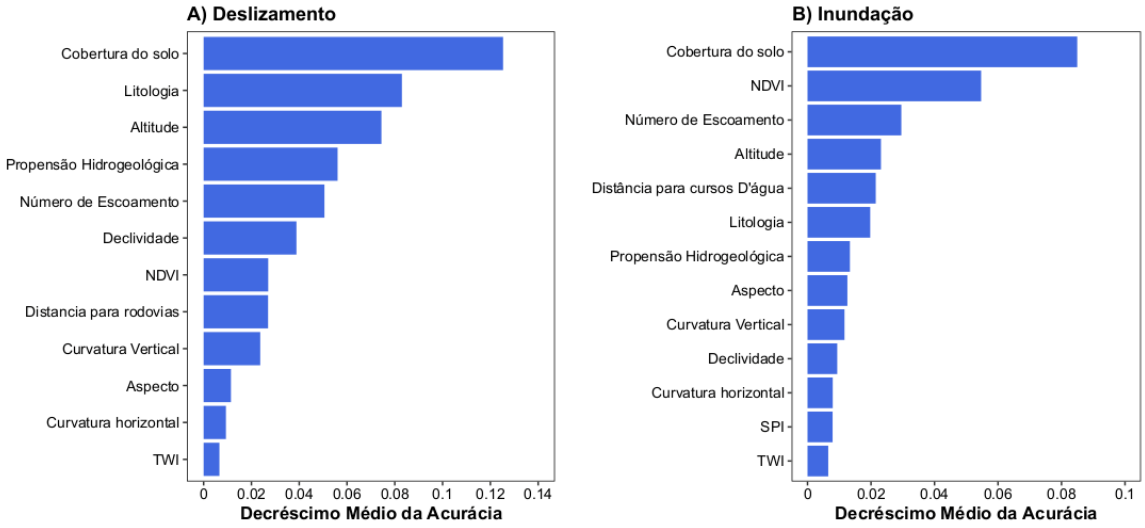
O modelo para classificação de deslizamento ficou 1,2 desvios padrões acima da média, quando analisado segundo a métrica AUC. Segundo a métrica da acurácia ficou 1,4 desvios padrões acima da média. O modelo de classificação de inundação ficou 0,4 desvios padrões acima da média segundo a métrica AUC. Já no critério acurácia ficou 0,6 desvios padrões abaixo.

Uma importante característica do *Random Forest* é a possibilidade de calcular a importância das variáveis, que serve para prover informações do valor preditivo de cada variável para ao modelo. No presente estudo, como discutido previamente, a métrica utilizada foi o decréscimo médio da acurácia, como exposto no Gráfico 16.

A variável com maior valor preditivo no modelo de deslizamento é a cobertura do solo, como observado, tal variável fornece informações sobre os padrões de cobertura e uso do solo, informações que são de extrema importância para a estabilidade do solo. A litologia, por sua vez ocupa o segundo lugar, e informa sobre a formação rochosa do solo, fator que influencia

diretamente na estabilidade do mesmo. Outras variáveis com valor preditivo alto foram a propensão hidrogeologica, número de escoamento e a declividade, pois são responsáveis por uma diminuição na acurácia de mais de 5%, caso fossem retiradas do modelo.

Gráfico 16– Importância das variáveis para a) deslizamento b) inundação



Fonte: O Autor (2019)

Diferente do que intuitivamente pode ser esperado, a variável declividade ocupou apenas a sexta posição em termos de importância preditiva. Esse resultado indica que terrenos inclinados por si não representam perigo, mas sim, uma combinação entre vários fatores, como a cobertura e uso do solo, litologia da área e características hidrológicas do solo. Assim sendo, avaliações a respeito do perigo de um determinado local devem levar em consideração tais interações, o que seria muito custoso através de modelos determinísticos. Felizmente, os algoritmos de aprendizado de máquina possuem a vantagem de identificação de tais relações de forma rápida e efetiva.

Para o modelo de inundação, a cobertura do solo também apresentou o maior valor preditivo, que como discutido anteriormente fornece informações relevantes a respeito dos padrões de uso e cobertura do solo. Em segundo lugar o NDVI, variável que também fornece informações sobre a vegetação. Outra variável com valor preditivo alto para o modelo foi a número de escoamento, tal variável informa sobre a capacidade de formação de escoamento superficial, logo, tem influência direta na formação de inundações.

A categoria de cobertura do solo com maior número de deslizamentos e inundações foi a infraestrutura urbana. Tal questão tem influência direta no risco ao qual a população está sujeita,

pois aumenta a exposição da população ao perigo. Além disso, maior dano financeiro será causado devido à perda de infraestrutura devido aos eventos desastrosos. Esse tipo de cobertura tem influência direta em outras características, como por exemplo, o número de escoamento. Ao observar a distribuição dos eventos, na Tabela 11, é possível perceber que para números de escoamento acima de 80 ocorreram a grande maioria dos eventos de deslizamentos e inundações. Tais valores são característicos de solos com alta capacidade de formação de escoamento superficial, característica observada em localidade com coberturas artificias, como cobertura asfáltica, por exemplo.

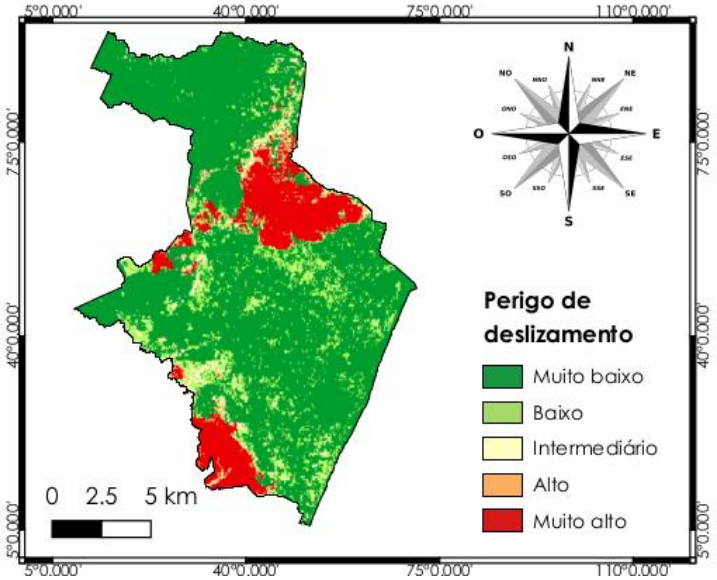
Um resultado que pode contrariar o senso de muitos é o de que a maior parte dos deslizamentos ocorreram em valores de número de escoamento acima de 80. Uma hipótese possível a respeito da relação entre o número de escoamento e o número de deslizamentos seria que quanto maior for o número de escoamento, maior é a probabilidade de formação de escoamento superficial e menor a capacidade de infiltração no solo, logo, se menos água é infiltrada no solo a estabilidade do mesmo seria maior. Porém, os resultados demonstraram o contrário, que mais deslizamentos ocorreram em altos números de escoamento. Uma possível explicação para isso é que como será formado escoamento nessas áreas, a estabilidade do solo será afetada devido ao poder erosivo do fluxo de água, formando crateras que favorecem os deslizamentos. De fato, ao classificar os deslizamentos em termos do SPI, foi possível observar que a grande maioria deles estão localizados em áreas com alto SPI.

O modelo ajustado obteve desempenho suficiente para ser colocado em produção. Dessa forma a próxima etapa consiste na produção dos mapas de perigo de deslizamento e inundações para a cidade de Recife-PE.

6.5 Elaboração do mapa de perigo e verificação do modelo

Para gerar os mapas de perigo de deslizamento e inundação foram utilizados dois recursos. O primeiro deles, como já citado, foi a linguagem de programação R (R CORE TEAM, 2019), com os pacotes citados anteriormente, só que o objetivo dessa etapa é realizar as previsões para pontos em toda a extensão da cidade de Recife. O segundo, software GIS *Qgis* (QGIS DEVELOPMENT TEAM, 2019), utilizado para gerar os pontos geograficamente distribuídos, bem como coletar os valores das variáveis para os novos pontos gerados. Após as previsões serem realizadas o arquivo foi então convertido para o formato Raster para dar origem ao mapa de perigo, conforme exibição do Mapa 5 e Mapa 6.

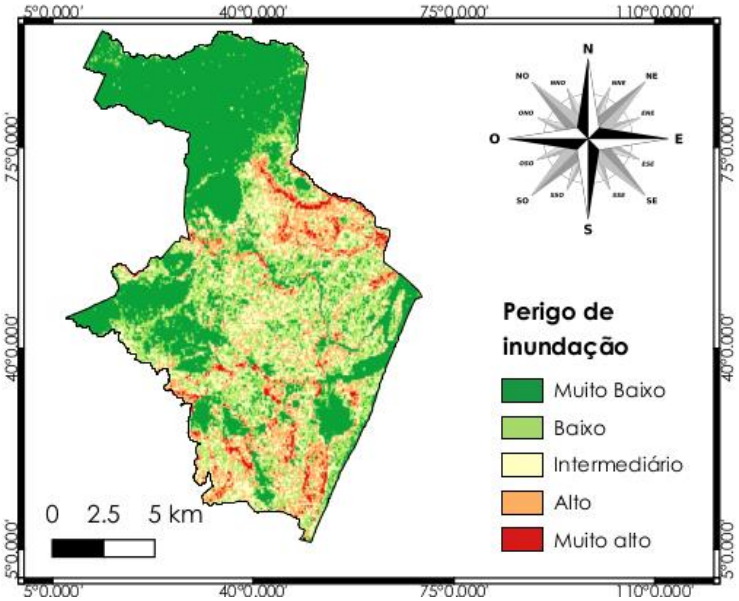
Mapa 5– Mapa de perigo para deslizamentos de terra



Fonte: O Autor (2019)

O mapa de perigo de deslizamento gerado pode ser caracterizado por 3 áreas bem definidas, com alto índice de perigo. Tais áreas são caracterizadas pela existência de morros e alto índice de moradias irregulares, o que aumenta a probabilidade de ocorrência de deslizamentos. Ainda no mapa é possível perceber áreas com índice de risco intermediário, locais nos quais prevalece a ocorrência de morros e fatores que aumentam a probabilidade do desastre.

Mapa 6– Mapa de perigo para inundações



Fonte: O Autor (2019)

O mapa de perigo para inundação e alagamentos possui índices mais uniformemente distribuídos. Um resultado interessante é que nas imediações dos rios, o índice de perigo é alto. Além disso, a região central da cidade, que é rodeada de curso d'água, possui índice de perigo intermediário e alto. Tal resultado indica que os moradores de regiões próximas a rios estão mais sujeitos a eventos de inundação e alagamento.

Apesar da validação da performance do modelo ter apresentado resultados satisfatórios, é importante realizar também validações qualitativas. No presente estudo, os resultados gerados através dos modelos de aprendizado de máquina foram comparados com dados disponíveis em notícias jornalísticas. Para isso, as notícias foram coletadas e extraídas informações do tipo do desastre e localização, conforme exposição do Quadro 9. As buscas foram realizadas através da internet com palavras chaves referentes a deslizamentos e inundações em jornais reconhecidos.

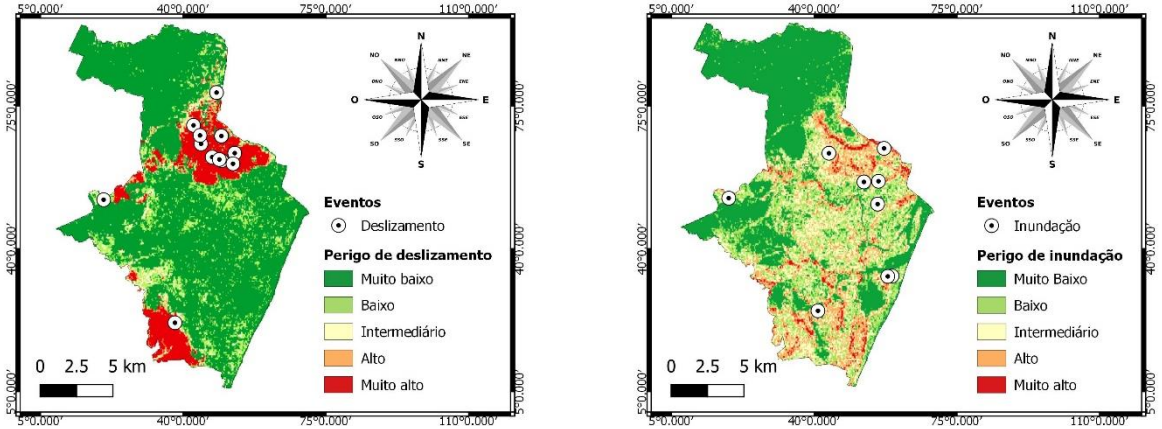
Quadro 9 – Desastres reportados em matérias jornalísticas

Desastre	**Bairro**	**Fonte**
Deslizamento	Dois Unidos	(G1 PE E TV GLOBO, 2019)
Deslizamento	Passarinho	(G1 PE E TV GLOBO, 2019)
Deslizamento	Dois Unidos	(G1 PE, 2019a)
Inundação	Avenida Antônio de Góes	(G1 PE, 2019a)
Inundação	canal de Setúbal	(G1 PE, 2019a)
Inundação	Avenida Nova Descoberta	(G1 PE, 2019a)
Inundação	Avenida Norte	(G1 PE, 2019a)
Inundação	Avenida Professor José dos Anjos	(G1 PE, 2019a)
Deslizamento	Linha do Tiro	(DIARIO DE PERNAMBUCO, 2019)
Inundação	Avenidas Norte	(G1 PE, 2019b)
Inundação	Conselheiro Portela	(G1 PE, 2019b)
Inundação	Várzea	(G1 PE, 2019b)
Deslizamento	Nova Descoberta	(G1 PE, 2018)
Inundação	Porto da Madeira	(G1 PE, 2018)
Deslizamento	Dois Unidos	(G1 PE, 2018)
Deslizamento	Vasco da Gama	(G1 PE, 2018)
Deslizamento	Linha do Tiro	(G1 PE, 2018)
Deslizamento	Dois Unidos	(G1 PE, 2018)
Deslizamento	Dois Unidos	(PORTAL FOLHAPE, 2018)
Deslizamento	Brejo da Guabiraba	(NE 10, 2016)
Deslizamento	Alto José Bonifácio	(JORNAL A TARDE, 2008)
Deslizamento	Córrego do Euclides	(CORREIO DO BRASIL, 2015)
Deslizamento	Várzea	(CORREIO DO BRASIL, 2015)
Deslizamento	Alto do Pascoal	(G1 PE, 2016)
Deslizamento	Guabiraba	(TV OUL, 2018)
Deslizamento	Ibura	(DIÁRIO DE PERNAMBUCO, 2015)

Fonte: O Autor (2019)

A localização de cada evento foi extraída conforme as informações disponíveis na matéria jornalística. Quando a localização exata do evento não estava disponível na notícia, e sim apenas o bairro, as coordenadas coletadas foram a do centroide do bairro em questão. Após a coleta, foi possível realizar a comparação entre o resultado produzido pelo modelo e os desastres reportados nas notícias jornalísticas, conforme exposto no Mapa 7. Os pontos representam os eventos coletados nas notícias.

Mapa 7– Validação do modelo com notícia jornalísticas



Fonte: O Autor (2019)

Como é possível observar no Mapa 7, todos os eventos coletados em notícia estão localizados em áreas com classe de perigo intermediária, alta ou muito alta. Os eventos de inundação estão localizados em área de alto perigo ou em áreas vizinhas. Já os eventos de deslizamento possuem correlação geográfica muito alta com os resultados produzidos pelo modelo.

Tal resultado significa que o modelo proposto foi capaz de mapear o perigo de desastres de forma muito similar ao que realmente acontece.

As avaliações da performance dos modelos foram realizadas em duas etapas, uma utilizando medidas estatísticas e a segunda, tão importante quanto, para verificar se os resultados produzidos foram condizentes com a realidade. Os resultados da primeira avaliação demonstraram que os modelos possuem desempenho igual ou superior aos modelos prospectados na literatura. Já a segunda avaliação demonstrou que os resultados produzidos são condizentes com o que de fato acontece, ou seja, as regiões classificadas com alto perigo realmente estão sujeitas aos desastres.

6.6 Discussões

A principal contribuição do presente trabalho e fator que o diferencia dos demais trabalhos avaliados durante a revisão da literatura conduzida foi a proposição de um modelo que viabiliza a utilização de dados semiestruturados na forma textual para o mapeamento de perigo de deslizamentos de terra e inundações. Tal modelo permite extrair informações de dados inutilizados até então para tal finalidade. Vale a pena ressaltar que o inventário construído possui escala regional com precisão de 30m, onde estão identificados eventos de deslizamentos de terra, inundações e situações potencialmente perigosas. Tal inventário pode ser utilizado para qualquer finalidade que necessite de informações sobre a localização dos eventos com precisão. Um outro ponto forte está no fato que o inventário é atualizado automaticamente, graças a criação do modelo de classificação textual. Dessa forma, novas informações sempre serão adicionadas, o que permite monitorar as mudanças e tendências na distribuição dos eventos.

Os motivos pelos quais é possível justificar a escolha do modelo proposto nesse trabalho pode ser dividida em quatro categorias: (*i*) custo; (*ii*) desempenho; (*iii*) simulação; (*iv*) integração.

Por se tratar de um modelo que utiliza aprendizado de máquina, tema bem discutido na literatura, e um algoritmo não-proprietário, a utilização do modelo proposto por parte do decisor permite não apenas uma redução significativa nos custos, uma vez que um especialista em modelos de aprendizado de máquina conseguirá compreender todas as etapas de aplicação e desenvolvimento, mas também possibilita a implementação de melhorias em aplicações futuras. Estas melhorias dizem respeito ao fato de que métodos de aprendizado de máquina podem ser incrementados através da utilização de mais dados, mais variáveis, engenharia de variáveis, diferentes algoritmos de aprendizado de máquina e utilização de heurísticas para otimização dos parâmetros do modelo. Soma-se a isso o fato de que uma característica marcante dos métodos de aprendizado de máquina é o de ser generalista, permitindo a aplicação, guardadas as devidas restrições conceituais, em cenários análogos. No caso do método desenvolvido em questão, o modelo pode ser aplicado para outros desastres naturais ou utilizado em outras localidades.

O desempenho do modelo proposto demonstrou ser suficiente para ser utilizado com segurança por decisores no processo de tomada de decisão. Tanto em termos de acurácia, ou seja, capacidade do modelo em aprender com o exemplo, bem como em velocidade de processamento. Isso proporciona informações seguras sobre o grau de perigo em questões de

minutos, permitindo avaliação e tomadas de decisões rápidas para o gerenciamento do risco de deslizamentos de terra e inundações.

Por utilizar fatores condicionantes e estabelecer uma relação explicativa com o grau de perigo, o modelo proposto pode também ser utilizada para simular cenários futuros através da inserção de ruídos controlados nos fatores condicionantes e verificando as modificações na variável resposta. Essa característica viabiliza o estudo da dinâmica dos eventos sob cenários de mudanças climáticas, alterações geomorfológicas, mudanças na infraestrutura urbana, testes de medidas de prevenção, *etc*.

Além disso, todas as funcionalidades descritas anteriormente são facilmente integradas com soluções tecnológicas já em utilização, sem necessariamente realizar grandes modificações de infraestrutura de tecnologia. Essa é uma das principais vantagens, pois sem grandes modificações estruturais, entidades responsáveis pelo gerenciamento do risco podem começar imediatamente a implantar a solução proposta, utilizando dados já existentes no banco de dados que antes não eram utilizados para tal fim.

Tais características provocam um impacto social, ao passo que a população será beneficiada com uma avaliação do perigo realizada de forma segura à um menor custo, tanto em termos monetários, computacionais e esforços humanos. Dessa forma, ao economizar recursos financeiros na identificação do grau de perigo em si, as entidades competentes podem remanejar tais recursos para o tratamento do risco, fornecendo respostas mais efetivas e rápidas para a população.

6.7 Conclusões do capítulo

No presente capítulo foi demonstrado sistematicamente a aplicação do modelo proposto para a cidade de Recife, situada no estado de Pernambuco, nordeste do Brasil. As variáveis utilizadas no modelo foram coletadas e tratadas para a área de estudo e seus valores descritos.

Após o treinamento do modelo, a avaliação de performance foi realizada tomando como base métricas advindas da validação cruzada e da curva ROC. Os modelos ajustados demostraram desempenho igual ou superior ao padrão estabelecido segundo os dados coletados na revisão sistemática da literatura, capítulo 4 do presente trabalho.

Apesar da abordagem proposta ter sido validada com um estudo de caso na cidade de Recife, a mesma pode ser aplicada em outras localidades, resguardando os devidos cuidados. Os dados semiestruturados devem possuir um conteúdo que relate o acontecimento de algum tipo de desastre para tornar possível a identificação por parte do algoritmo de classificação de

texto. Além disso, as variáveis condicionantes utilizadas devem ser escolhidas levando em consideração o tipo de desastre estudado e as características da área de estudo.

Os mapas de perigo para inundações e deslizamentos foram criados, uma vez que o desempenho foi suficiente. Tais mapas foram comparados com inundações e deslizamentos reportados em notícias jornalísticas. O modelo demonstrou ser capaz de descrever a realidade uma vez que os eventos coletados nas matérias jornalísticas estavam localizados em áreas classificadas como perigosa pelos modelos ajustados.

Por fim, como todos os testes realizados com o modelo retornaram resultados positivos, o modelo pode ser posto em produção, auxiliando na tomada de decisão no gerenciamento de risco de desastres naturais, observando os aspectos discutidos na seção 6.6.

7 CONCLUSÕES E TRABALHOS FUTUROS

O presente estudo teve como objetivo geral propor um modelo para mapeamento de perigo de deslizamentos de terra e inundações, utilizando dados semiestruturados advindos de linguagem natural na forma textual. Com esse modelo formar um inventário de eventos georreferenciados para ser utilizado no mapeamento do perigo de deslizamentos de terra e inundações, utilizando algoritmos de aprendizado de máquina.

Foi estabelecido um processo para mapeamento de perigo de desastres naturais baseados em algoritmos de aprendizado de máquina, baseado na revisão sistemática da literatura. Além disso, foi estabelecido um conjunto de informações úteis para a comparação de novos modelos com modelos previamente publicados na literatura.

O modelo de tratamento e classificação textual foi ajustado e obteve resultados satisfatórios. Ambos os modelos obtiveram desempenho igual ou superior ao padrão estabelecido na revisão sistemática da literatura. Os mapas de perigo foram comparados com eventos coletados através de notícias jornalísticas, mostrando novamente que os modelos são capazes de descrever a dinâmicos de ocorrência dos eventos. Com base nesses resultados, é possível afirmar que tanto o objetivo geral como os objetivos específicos do trabalho foram atendidos de forma satisfatória e sistematicamente apresentados.

O modelo desenvolvido no presente estudo apresentou duas principais limitações. A primeira delas é em relação a possibilidade de extrapolação dos dados para áreas fora dos limites da área de estudo. Para isso, ainda será necessário testar a abordagem em outras áreas para verificar a possibilidade de reprodutibilidade dos resultados. A segunda limitação é a impossibilidade, segundo a estrutura atual, de indexar temporalmente os dados de precipitação pluviométrica para verificar a influência de tal variável nos resultados do perigo de inundações e deslizamentos.

Um ponto importante a ser notado no modelo proposto é que na forma como foi entregue, cumpre muito bem o papel descritivo do perigo, ou seja, consegue identificar e descrever o grau de perigo que cada ponto está sujeito com base em exemplos e variáveis atuais. Porém, tal modelo ainda não capaz de executar a tarefa de predição do perigo em determinado período de tempo futuro, inviabilizando estudo sob a perspectiva de mudanças climáticas.

Uma das dificuldades encontradas para a modelagem dos chamados foi a grande quantidade de erros ortográficos, aumentando significativamente a quantidade de termos a serem analisados pelo modelo, o que diminui tanto a acurácia do modelo como a performance computacional. Assim sendo, como melhoria para o modelo de classificação textual

recomenda-se a elaboração de um algoritmo de correção ortográfica para diminuir o tamanho do espaço das variáveis, aumentando assim a acurácia e o desempenho computacional.

Para o ajuste dos modelos de mapeamento de perigo de desastre, recomenda-se a adição de uma etapa anterior ao treinamento para a seleção das variáveis com maior valor preditivo. Tal etapa irá reduzir o custo computacional do modelo, pois apenas as variáveis mais significativas serão adicionadas.

Outra sugestão de melhoria no modelo é a adição da precipitação pluviométrica como uma variável no modelo. Para a adição de tal variável, é necessário realizar modificações estruturais na forma como o modelo foi construído, possibilitando o cruzamento entre os pontos que representam os desastres e a precipitação equivalente naquele dia.

Um ponto discutido na literatura específica de gestão de riscos é que a avaliação do risco não deve ser baseada totalmente em análises numéricas. As informações qualitativas advindas de decisores experientes também devem ser incorporadas no processo de gerenciamento do risco. Assim sendo, um paradigma de aprendizado de máquina que deve ser estudado com maior profundidade é o aprendizado simbólico, pois este permite ao decisor identificar, compreender e julgar adequada ou não a estrutura lógica de decisão definida pelo algoritmo. Por se tratar de um tema no qual os danos ultrapassam os financeiros, lidando diretamente com vidas humanas, os algoritmos conhecidos como “caixa preta” podem oferecer maior dificuldade no entendimento e comunição por parte dos decisores, dificultando o processo de análise de risco.

Isto posto, conclui-se que o modelo proposto no presente trabalho é capaz de realizar o mapeamento do perigo de deslizamentos e inundação utilizando dados semiestruturados advindos de linguagem natural na forma textual, utilizando o algoritmo *Naive Bayes* para a classificação textual e o algoritmo *Random Forest* para o mapeamento do perigo de deslizamentos e inundações.

REFERÊNCIAS

ADA, M.; SAN, B. T. Comparison of machine-learning techniques for landslide susceptibility mapping using two-level random sampling (2LRS) in Alakir catchment area, Antalya, Turkey. *Natural Hazards*, v. 90, n. 1, p. 237–263, 2018.

ADINEH, F. *et al*. Landslide susceptibility mapping using Genetic Algorithm for the Rule Set Production (GARP) model. *Journal of Mountain Science*, v. 15, n. 9, p. 2013–2026, 2018.

AKGUN, A.; KINCAL, C.; PRADHAN, B. Application of remote sensing data and GIS for landslide risk assessment as an environmental threat to Izmir city (west Turkey). *Environmental Monitoring and Assessment*, v. 184, n. 9, p. 5453–5470, 2012.

AL-ABADI, A. M. Mapping flood susceptibility in an arid region of southern Iraq using ensemble machine learning classifiers: a comparative study. *Arabian Journal of Geosciences*, v. 11, n. 9, 2018.

ALEXANDER, M.; PRIEST, S.; MEES, H. A framework for evaluating flood risk governance. *Environmental Science and Policy*, v. 64, p. 38–47, 2016.

ALMEIDA, A. T. DE *et al*. **Multicriteria and Multiobjetive Models for Risk, Reliability and Maintenance Decision Analysis**. International Series in Operations Research & Management Science. **Anais**...New York: Springer, 2015

ALTHUWAYNEE, O. F. *et al*. A novel ensemble decision tree-based CHi-squared Automatic Interaction Detection (CHAID) and multivariate logistic regression models in landslide susceptibility mapping. *Landslides*, v. 11, n. 6, p. 1063–1078, 2014.

ALVES, A. *et al*. Multi-criteria Approach for Selection of Green and Grey Infrastructure to Reduce Flood Risk and Increase CO-benefits. *Water Resources Management*, v. 32, n. 7, p. 2505–2522, 2018.

ANA, A. N. DE Á. **GeoNEtwork**. Disponível em: <https://metadados.ana.gov.br/geonetwork/srv/pt/main.home>. Acesso em: 5 fev. 2016.

APURV, T. *et al*. Impact of climate change on floods in the Brahmaputra basin using CMIP5 decadal predictions. *Journal of Hydrology*, v. 527, p. 281–291, 2015.

ARABAMERI, A. *et al*. A comparison of statistical methods and multi-criteria decision making to map flood hazard susceptibility in Northern Iran. *Science of the Total Environment*, v. 660, p. 443–458, 2019.

ARABAMERI, A.; POURGHASEMI, H. R.; YAMANI, M. Applying different scenarios for landslide spatial modeling using computational intelligence methods. *Environmental Earth Sciences*, v. 76, n. 24, 2017.

ARIA, M.; CUCCURULLO, C. bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, v. 11, n. 4, p. 959–975, 2017.

ARNONE, E. *et al*. Strategies investigation in using artificial neural network for landslide susceptibility mapping: application to a Sicilian catchment. *Journal of Hydroinformatics*, v. 16, n. 2, p. 502–515, 2014.

ARTISSA, Y. B. N. D.; ASROR, I.; FARABY, S. A. **Personality Classification based on Facebook status text using Multinomial Naïve Bayes method**. Journal of Physics: Conference Series. **Anais**...2019

AVEN, T. **Risk Analysis**. 2. ed. Chichester: John Wiley & Sons, 2015.

AVEN, T.; RENIERS, G. How to define and interpret a probability in a risk and safety setting. *Safety Science*, v. 51, n. 1, p. 223–231, 2013.

BALAMURUGAN, G.; RAMESH, V.; TOUTHANG, M. Landslide susceptibility zonation mapping using frequency ratio and fuzzy gamma operator models in part of NH-39, Manipur, India. *Natural Hazards*, v. 84, n. 1, p. 465–488, 2016.

BALLABIO, C.; STERLACCHINI, S. Support Vector Machines for Landslide Susceptibility Mapping: The Staffora River Basin Case Study, Italy. *Mathematical Geosciences*, v. 44, n. 1, p. 47–70, 2012.

BEHNIA, P.; BLAIS-STEVENS, A. Landslide susceptibility modelling using the quantitative random forest method along the northern portion of the Yukon Alaska Highway Corridor, Canada. *Natural Hazards*, v. 90, n. 3, p. 1407–1426, 2018.

BEN-DAVID, S.; SHALEV-SHWARTZ, S. **Understanding Machine Learning: From Theory to Algorithms**. [s.l: s.n.].

BEVEN, K. J.; KIRKBY, M. J. A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, v. 24, n. 1, p. 43–69, 1979.

BORNAETXEA, T. *et al*. Effective surveyed area and its role in statistical landslide susceptibility assessments. *Natural Hazards and Earth System Sciences*, v. 18, n. 9, p. 2455–2469, 2018.

BRASIL. **Manual de Proteção e Defesa Civil: Glossário de Proteção e Defesa Civil**. 5ª ed. Brasília: Secretaria Nacional de Proteção e Defesa Civil – SEDEC, 2009.

BRASIL. **Atlas Brasileiro de Desastres Naturais: 1991 a 2012**. 2ª ed. Florianópolis: CEPED UFSC, 2013.

BREIMAN, L. Random Forest. *Machine Learning*, v. 45, p. 5–32, 2001.

BRERETON, P. *et al*. Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, v. 80, n. 4, p. 571–583, 2007.

BUI, D. T. *et al*. Landslide susceptibility analysis in the Hoa Binh province of Vietnam using statistical index and logistic regression. *Natural Hazards*, v. 59, n. 3, p. 1413–1444, 2011.

BUI, D. T. *et al*. Landslide detection and susceptibility mapping by AIRSAR data using support vector machine and index of entropy models in Cameron Highlands, Malaysia. *Remote Sensing*, v. 10, n. 10, 2018.

CALLE, M. L.; URREA, V. Letter to the editor: Stability of Random Forest importance measures. *Briefings in Bioinformatics*, v. 12, n. 1, p. 86–89, 2011.

CAMPBELL, S. Determining overall risk. *Journal of Risk Research*, v. 8, n. 7–8, p. 569–581, 2005.

CASTRO, C. L. DE; BRAGA, A. P. Aprendizado supervisionado com conjuntos de dados desbalanceados. *Sba: Controle & Automação Sociedade Brasileira de Automatica*, v. 22, n. 5, p. 441–466, 2012.

CERQUEIRA, D. *et al*. **Atla da Violência 2019 - Retratos dos municípios brasileiros**. Rio de Janeiro: IPEA - Instituto de Pesquisa Econômica e Aplicada, 2019.

CHANG, L. C. *et al*. Clustering-based hybrid inundation model for forecasting flood inundation depths. *Journal of Hydrology*, v. 385, n. 1–4, p. 257–268, 2010.

CHEN, W. *et al*. Landslide susceptibility mapping based on GIS and information value model for the Chencang District of Baoji, China. *Arabian Journal of Geosciences*, v. 7, n. 11, p. 4499–4511, 2014.

CHEN, W. *et al*. Spatial prediction of landslide susceptibility using integrated frequency ratio with entropy and support vector machines by different kernel functions. *Environmental Earth Sciences*, v. 75, n. 20, 2016.

CHEN, W. *et al*. GIS-based landslide susceptibility modelling: a comparative assessment of kernel logistic regression, Naïve-Bayes tree, and alternating decision tree models. *Geomatics, Natural Hazards and Risk*, v. 8, n. 2, p. 950–973, 2017a.

CHEN, W. *et al*. Spatial prediction of landslide susceptibility using an adaptive neuro-fuzzy inference system combined with frequency ratio, generalized additive model, and support vector machine techniques. *Geomorphology*, v. 297, p. 69–85, 2017b.

CHEN, W. *et al*. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena*, v. 151, p. 147–160, 2017c.

CHEN, W. *et al*. Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling. *Science of the Total Environment*, v. 644, p. 1006–1018, 2018a.

CHEN, W. *et al*. Landslide susceptibility modelling using GIS-based machine learning techniques for Chongren County, Jiangxi Province, China. *Science of the Total Environment*, v. 626, p. 1121–1135, 2018b.

CHEN, W. *et al*. GIS-based landslide susceptibility evaluation using a novel hybrid integration approach of bivariate statistical based random forest method. *Catena*, v. 164, n. April 2017, p. 135–149, 2018c.

CHEN, W.; POURGHASEMI, H. R.; NAGHIBI, S. A. Prioritization of landslide conditioning factors and its spatial modeling in Shangnan County, China using GIS-based data mining algorithms. *Bulletin of Engineering Geology and the Environment*, v. 77, n. 2, p. 611–629, 2018a.

CHEN, W.; POURGHASEMI, H. R.; NAGHIBI, S. A. A comparative study of landslide susceptibility maps produced using support vector machine with different kernel functions and entropy data mining models in China. *Bulletin of Engineering Geology and the Environment*, v. 77, n. 2, p. 647–664, 2018b.

CONFORTI, M. *et al*. Evaluation of prediction capability of the artificial neural networks for mapping landslide susceptibility in the Turbolo River catchment (northern Calabria, Italy). *Catena*, v. 113, p. 236–250, 2014.

CORREIO DO BRASIL. **Recife: homem desaparece em deslizamento causado pela chuva**. Disponível em: <https://arquivo.correiodobrasil.com.br/recife-homem-desaparece-em-deslizamento-causado-pela-chuva/>. Acesso em: 30 out. 2019.

COSTANZO, D. *et al*. Forward logistic regression for earth-flow landslide susceptibility assessment in the Platani river basin (southern Sicily, Italy). *Landslides*, v. 11, n. 4, p. 639–653, 2014.

CPRM, S. G. DO B. **GeoSGB: Dados, informações e produtos do Serviço Geológico do Brasil**. Disponível em: <http://geosgb.cprm.gov.br>. Acesso em: 5 fev. 2019.

DAHAL, R. K. Regional-scale landslide activity and landslide susceptibility zonation in the Nepal Himalaya. *Environmental Earth Sciences*, v. 71, n. 12, p. 5145–5164, 2014.

DE, U. S.; KHOLE, M.; DANDEKAR, M. M. Natural hazards associated with meteorological extreme events. *Natural Hazards*, v. 31, n. 2, p. 487–497, 2004.

DELEO, J. M. **Receiver operating characteristic laboratory (ROCLAB): Software for developing decision strategies that account for uncertainty**. International Symposium on Uncertainty Modeling and Analysis. **Anais**...College Park, MD, USA, USA: IEEE, 1993

DEMIR, G. *et al*. A comparison of landslide susceptibility mapping of the eastern part of the North Anatolian Fault Zone (Turkey) by likelihood-frequency ratio and analytic hierarchy process methods. *Natural Hazards*, v. 65, n. 3, p. 1481–1506, 2013.

DIARIO DE PERNAMBUCO. **Deslizamento de terra causa acidente no Recife**. Disponível em: <https://www.diariodepernambuco.com.br/noticia/vidaurbana/2019/09/deslizamento-de-terra-causa-acidente-no-recife.html>. Acesso em: 28 out. 2019.

DIÁRIO DE PERNAMBUCO. **Barreira desliza no Ibura deixa três feridos e atinge quatro casas**. Disponível em: <https://www.diariodepernambuco.com.br/noticia/vidaurbana/2015/03/barreira-desliza-no-ibura-deixa-tres-feridos-e-atinge-quatro-casas.html%0A>. Acesso em: 30 out. 2019.

EM-DAT. **The Emergency Events Database**. Disponível em: <www.emdat.be>. Acesso em: 1 jan. 2019.

ERCANOGLU, M.; TEMIZ, F. A. Application of logistic regression and fuzzy operators to landslide susceptibility assessment in Azdavay (Kastamonu, Turkey). *Environmental Earth Sciences*, v. 64, n. 4, p. 949–964, 2011.

ERMINI, L.; CATANI, F.; CASAGLI, N. Artificial Neural Networks applied to landslide susceptibility assessment. *Geomorphology*, v. 66, n. 1- 4 SPEC. ISS., p. 327–343, 2005.

FARAJI SABOKBAR, H.; SHADMAN ROODPOSHTI, M.; TAZIK, E. Landslide susceptibility mapping using geographically-weighted principal component analysis. *Geomorphology*, v. 226, p. 15–24, 2014.

FEIZIZADEH, B. *et al*. Comparing GIS-based support vector machine kernel functions for landslide susceptibility mapping. *Arabian Journal of Geosciences*, v. 10, n. 5, 2017.

FELICÍSIMO, Á. M. *et al*. Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: A comparative study. *Landslides*, v. 10, n. 2, p. 175–189, 2013.

FENG, H. *et al*. Evaluation of different models in rainfall-triggered landslide susceptibility mapping: a case study in Chunan, southeast China. *Environmental Earth Sciences*, v. 75, n. 21, 2016.

FENG, Q. *et al*. Flood mapping based on multiple endmember spectral mixture analysis and random forest classifier-the case of yuyao, China. *Remote Sensing*, v. 7, n. 9, p. 12539–12562, 2015.

FENG, Q.; LIU, J.; GONG, J. Urban flood mapping based on unmanned aerial vehicle remote sensing and random forest classifier-A case of yuyao, China. *Water (Switzerland)*, v. 7, n. 4, p. 1437–1455, 2015.

FIELDING, A. H.; BELL, J. F. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, v. 24, n. 1, p. 38–49, 1997.

FLORINSKY, I. V (ED.). **Digital Terrain Analysis in Soil Science and Geology**. Boston: Academic Press, 2012.

FRANK, E.; BOUCKAERT, R. R. Naive bayes for text classification with unbalanced classes. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 4213 LNAI, p. 503–510, 2006.

FRIGERIO, I. *et al*. A GIS-based approach to identify the spatial variability of social vulnerability to seismic hazard in Italy. *Applied Geography*, v. 74, p. 12–22, 2016.

G1 PE. **Barreira desliza no Alto do Pascoal, Zona Norte do Recife, e deixa uma casa destruída**. Disponível em: <http://g1.globo.com/pernambuco/videos/v/barreira-desliza-no-alto-do-pascoal-zona-norte-do-recife-e-deixa-uma-casa-destruida/5007191/>. Acesso em: 30 out. 2019.

G1 PE. **Chuva provoca deslizamentos de barreiras e alagamento de ruas no Recife**. Disponível em: <https://g1.globo.com/pe/pernambuco/noticia/chuva-provoca-deslizamentos-de-barreiras-e-alagamento-de-ruas-no-recife.ghtml>. Acesso em: 28 out. 2019.

G1 PE. **Chuva alaga ruas, provoca desabamento de parte de casa e suspende aulas no Grande Recife**. Disponível em: <https://g1.globo.com/pe/pernambuco/noticia/2019/06/13/chuva-provoca-deslizamento-e-alagamentos-no-recife.ghtml>. Acesso em: 28 out. 2019a.

G1 PE. **Chuva alaga ruas e causa transtornos no Grande Recife**. Disponível em: <https://g1.globo.com/pe/pernambuco/noticia/2019/04/12/chuva-alaga-ruas-e-causa-transtornos-no-recife.ghtml>. Acesso em: 28 out. 2019b.

G1 PE E TV GLOBO. **Chuva causa mortes, deslizamento de barreiras e alagamentos no Grande Recife**. Disponível em:

<https://g1.globo.com/pe/pernambuco/noticia/2019/07/24/chuva-causa-deslizamento-de-barreiras-e-alagamento-no-grande-recife.ghtml>. Acesso em: 28 out. 2019.

GAIDZIK, K. *et al*. Landslide manual and automated inventories, and susceptibility mapping using LIDAR in the forested mountains of Guerrero, Mexico. *Geomatics, Natural Hazards and Risk*, v. 8, n. 2, p. 1054–1079, 2017.

GARCÍA-RODRÍGUEZ, M. J. *et al*. Susceptibility assessment of earthquake-triggered landslides in El Salvador using logistic regression. *Geomorphology*, v. 95, n. 3–4, p. 172–191, 2008.

GARCÍA-RODRÍGUEZ, M. J.; MALPICA, J. A. Assessment of earthquake-triggered landslide susceptibility in El Salvador based on an artificial neural network model. *Natural Hazards and Earth System Science*, v. 10, n. 6, p. 1307–1315, 2010.

GARRICK, J.; KAPLAN, S.; HAIMES, Y. Fitting Hierarchical Holographic Modeling into the Theory of Scenario Structuring and a Resulting Refinement to the Quantitative Definition of Risk. *Risk Analysis*, v. 21, n. 5, p. 807–807, 2001.

GHEORGHIU, A. D. *et al*. COMPARATIVE ANALYSIS OF TECHNOLOGICAL AND NATECH RISK FOR TWO PETROLEUM PRODUCT TANKS LOCATED IN SEISMIC AREA. *Environmental Engineering and Management Journal*, v. 13, n. 8, p. 1887–1892, 2014.

GIL, A. C. **Como elaborar projetos de pesuisa**. Pioneira ed. São Paulo: [s.n.].

GIOVANNETTONE, J. *et al*. A Statistical Approach to Mapping Flood Susceptibility in the Lower Connecticut River Valley Region. *Water Resources Research*, v. 54, n. 10, p. 7603–7618, 2018.

GISLASON, P. O.; BENEDIKTSSON, J. A.; SVEINSSON, J. R. Random forests for land cover classification. *Pattern Recognition Letters*, v. 27, n. 4, p. 294–300, 2006.

GOETZ, J. N. *et al*. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Computers and Geosciences*, v. 81, p. 1–11, 2015.

GOKCEOGLU, C. *et al*. Assessment of landslide susceptibility by decision trees in the metropolitan area of Istanbul, Turkey. *Mathematical Problems in Engineering*, v. 2010, 2010.

GÓMEZ, H.; KAVZOGLU, T. Assessment of shallow landslide susceptibility using artificial neural networks in Jabonosa River Basin, Venezuela. *Engineering Geology*, v. 78, n. 1–2, p. 11–27, 2005.

GOSWAMI, S. *et al*. A Review on Application of Data Mining Techniques to Combat Natural Disasters. *Ain Shams Engineering Journal*, v. 9, n. 1, p. 365–378, 2018.

GUPTA, S. *et al*. Systematic Review of the Literature: Best Practices. *Academic Radiology*, v. 25, n. 11, p. 1481–1490, 2018.

GURI, P. K.; CHAMPATI RAY, P. K.; PATEL, R. C. Spatial prediction of landslide susceptibility in parts of Garhwal Himalaya, India, using the weight of evidence modelling. *Environmental Monitoring and Assessment*, v. 187, n. 6, 2015.

HA, K.-M. Learning or Ignoring Lessons from Natural Disasters. *Journal of Professional Issues in Engineering Education and Practice*, v. 145, n. 4, p. 02519001, 2019.

HÄBERLE, M.; WERNER, M.; ZHU, X. X. Geo-spatial text-mining from Twitter–a feature space analysis with a view toward building classification in urban regions. *European Journal of Remote Sensing*, v. 52, n. 2, p. 2–11, 2019.

HADI, W. EMCAR: Expert Multi Class Based on Association Rule. *International Journal of Modern Education and Computer Science*, v. 5, n. 3, p. 33–41, 2013.

HADI, W.; AL-RADAIDEH, Q. A.; ALHAWARI, S. Integrating associative rule-based classification with Naïve Bayes for text classification. *Applied Soft Computing Journal*, v. 69, p. 344–356, 2018.

HAN, X.; KWOH, C. K. Natural Language Processing Approaches in Bioinformatics. *Encyclopedia of Bioinformatics and Computational Biology*, v. 1, p. 561–574, 2019.

HARRIS, J. R. *et al*. Data- and knowledge-driven mineral prospectivity maps for Canada's North. *Ore Geology Reviews*, v. 71, p. 788–803, 2015.

HARTMANN, J. *et al*. Comparing automated text classification methods. *International Journal of Research in Marketing*, v. 36, n. 1, p. 20–38, 2019.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data mining, Inference, and Prediction**. 2ª ed. New York: SPRINGER, 2009.

HIGHLAND, L. M.; BOBROWSKY, P. **The Landslide Handbook— A Guide to Understanding Landslides**. Reston, Virginia: U.S. Geological Survey Circular 1325, 2008.

HOANG, N. D.; TIEN BUI, D. Spatial prediction of rainfall-induced shallow landslides using gene expression programming integrated with GIS: a case study in Vietnam. *Natural Hazards*, v. 92, n. 3, p. 1871–1887, 2018.

HONG, H. *et al*. Spatial prediction of landslide hazard at the Yihuang area (China) using two-class kernel logistic regression, alternating decision tree and support vector machines. *Catena*, v. 133, p. 266–281, 2015.

HONG, H. *et al*. Spatial prediction of landslide hazard at the Luxi area (China) using support vector machines. *Environmental Earth Sciences*, v. 75, n. 1, p. 1–14, 2016.

HONG, H. *et al*. Comparison of four kernel functions used in support vector machines for landslide susceptibility mapping: a case study at Suichuan area (China). *Geomatics, Natural Hazards and Risk*, v. 8, n. 2, p. 544–569, 2017a.

HONG, H. *et al*. A novel hybrid integration model using support vector machines and random subspace for weather-triggered landslide susceptibility assessment in the Wuning area (China). *Environmental Earth Sciences*, v. 76, n. 19, p. 1–19, 2017b.

HONG, H. *et al*. Application of fuzzy weight of evidence and data mining techniques in construction of flood susceptibility map of Poyang County, China. *Science of the Total Environment*, v. 625, p. 575–588, 2018.

HONG, H.; POURGHASEMI, H. R.; POURTAGHI, Z. S. Landslide susceptibility assessment in Lianhua County (China): A comparison between a random forest data mining technique and bivariate and multivariate statistical models. *Geomorphology*, v. 259, p. 105–118, 2016.

HUANG, Y.; ZHAO, L. Review on landslide susceptibility mapping using support vector machines. *Catena*, v. 165, n. January, p. 520–529, 2018.

INPE. **TOPODATA: Banco de dados geomorfológicos do Brasil**. Disponível em: <http://www.dsr.inpe.br/topodata/index.php>. Acesso em: 6 fev. 2019.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. **IBGE Cidades**. Disponível em: <https://cidades.ibge.gov.br/brasil/pe/recife>. Acesso em: 27 set. 2019.

INTARAWICHIAN, N.; DASANANDA, S. Frequency ratio model based landslide susceptibility mapping in lower Mae Chaem watershed, Northern Thailand. *Environmental Earth Sciences*, v. 64, n. 8, p. 2271–2285, 2011.

IPCC. **Managing the risks of extreme events and disasters to advance climate change adaptation**. Cambridge, UK, and New York, NY, USA: Cambridge University Press, 2012.

JABBARI, A.; BAE, D. H. Application of Artificial Neural Networks for accuracy enhancements of real-time flood forecasting in the Imjin basin. *Water (Switzerland)*, v. 10, n. 11, 2018.

JI, Z. *et al*. Comprehensive assessment of flood risk using the classification and regression tree method. *Stochastic Environmental Research and Risk Assessment*, v. 27, n. 8, p. 1815–1828, 2013.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. *Science*, v. 349, n. 6245, p. 255–260, 2015.

JORNAL A TARDE. **Idosa morre em deslizamento de terra no Recife**. Disponível em: <https://www.atarde.uol.com.br/brasil/noticias/1193505-idosa-morre-em-deslizamento-de-terra-no-recife>. Acesso em: 30 out. 2019.

KALANTAR, B. *et al*. Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN). *Geomatics, Natural Hazards and Risk*, v. 9, n. 1, p. 49–69, 2018.

KAVZOGLU, T.; SAHIN, E. K.; COLKESEN, I. Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. *Landslides*, v. 11, n. 3, p. 425–439, 2014.

KAYASTHA, P.; DHITAL, M. R.; DE SMEDT, F. Evaluation of the consistency of landslide susceptibility mapping: A case study from the Kankai watershed in east Nepal. *Landslides*, v. 10, n. 6, p. 785–799, 2013.

KHOSRAVI, K. *et al*. A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. *Science of the Total Environment*, v. 627, p. 744–755, 2018.

KIM, J. C. *et al*. Landslide susceptibility mapping using random forest and boosted tree models in Pyeong-Chang, Korea. *Geocarto International*, v. 33, n. 9, p. 1000–1015, 2018.

KINCAL, C.; AKGUN, A.; KOCA, M. Y. Landslide susceptibility assessment in the İzmir (West Anatolia, Turkey) city center and its near vicinity by the logistic regression method.

Environmental Earth Sciences, v. 59, n. 4, p. 745–756, 2009.

KITCHENHAM, B. *et al*. **Guidelines for performing Systematic Literature Reviews in Software Engineering**. 2.3 ed. UK: Keele University and Durham University: EBSE Technical Report, 2007. v. 2

KOURGIALAS, N. N.; KARATZAS, G. P. A national scale flood hazard mapping methodology: The case of Greece – Protection and adaptation policy approaches. *Science of the Total Environment*, v. 601–602, p. 441–452, 2017.

KUBAL, C. *et al*. Integrated urban flood risk assessment – adapting a multicriteria approach to a city. *Natural Hazards and Earth System Sciences*, v. 9, p. 1881–1895, 2009.

LAI, C. *et al*. Flood risk zoning using a rule mining based on ant colony algorithm. *Journal of Hydrology*, v. 542, p. 268–280, 2016.

LANDIS, J. R.; KOCH, G. G. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, v. 33, n. 1, p. 159–174, 1977.

LANTZ, B. **Machine learning with R**. 2ª ed. Birmingham: Packt, 2015.

LEE, J. H. *et al*. Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods. *Geomorphology*, v. 303, p. 284–298, 2018.

LEE, S. *et al*. Use of an artificial neural network for analysis of the susceptibility to landslides at Boun, Korea. *Environmental Geology*, v. 44, n. 7, p. 820–833, 2003.

LEE, S. Landslide susceptibility mapping using an artificial neural network in the Gangneung are, Korea. *International Journal of Remote Sensing*, v. 28, n. 21, p. 4763–4783, 2007.

LEE, S. *et al*. Spatial Landslide Hazard Prediction Using Rainfall Probability and a Logistic Regression Model. *Mathematical Geosciences*, v. 47, n. 5, p. 565–589, 2015.

LEE, S. *et al*. Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. *Geomatics, Natural Hazards and Risk*, v. 8, n. 2, p. 1185–1203, 2017.

LEE, S.; HWANG, J.; PARK, I. Application of data-driven evidential belief functions to landslide susceptibility mapping in Jinbu, Korea. *Catena*, v. 100, p. 15–30, 2013.

LEE, S.; LEE, M. J.; LEE, S. Spatial prediction of urban landslide susceptibility based on topographic factors using boosted trees. *Environmental Earth Sciences*, v. 77, n. 18, p. 0, 2018.

LEE, S.; OH, H. J. Ensemble-based landslide susceptibility maps in jinbu area, Korea. *Terrigenous Mass Movements: Detection, Modelling, Early Warning and Mitigation Using Geoinformation Technology*, v. 9783642254, p. 193–220, 2014.

LIAW, A.; WIENER, M. Classification and Regression by randomForest. *R News*, v. 2, n. 3, p. 18–22, 2002.

LIU, X.; MIAO, C. Large-scale assessment of Landslide Hazard, vulnerability and risk in China. *Geomatics, Natural Hazards and Risk*, v. 9, n. 1, p. 1037–1052, 2018.

LOMBARDO, L. *et al*. Binary logistic regression versus stochastic gradient boosted decision trees in assessing landslide susceptibility for multiple-occurring landslide events: application to the 2009 storm event in Messina (Sicily, southern Italy). *Natural Hazards*, v. 79, n. 3, p. 1621–1648, 2015.

MACHADO JUNIOR, C. *et al*. Laws of Bibliometrics in Different Scientific Databases. *Revista de Ciências da Administração*, v. 18, n. 44, p. 111–123, 2016.

MAGANTI, N. *et al*. Natural Language Processing to Quantify Microbial Keratitis Measurements. *Ophthalmology*, p. 1–3, 2019.

MAPBIOMAS. **Collection 4 of Brazilian Land Cover & Use Map Series**. Disponível em: <http://mapbiomas.org>. Acesso em: 5 fev. 2019.

MELCHIORRE, C. *et al*. Evaluation of prediction capability, robustness, and sensitivity in non-linear landslide susceptibility models, Guantánamo, Cuba. *Computers and Geosciences*, v. 37, n. 4, p. 410–425, 2011.

MERGHADI, A.; ABDERRAHMANE, B.; TIEN BUI, D. Landslide Susceptibility Assessment at Mila Basin (Algeria): A Comparative Assessment of Prediction Capability of Advanced Machine Learning Methods. *ISPRS International Journal of Geo-Information*, v. 7, n. 7, p. 268, 2018.

MIGUEZ, M. G.; GREGORIO, L. T. DI; VERÓL, A. P. **Riscos e Desastres Hidrológicos**. 1. ed. Rio de Janeiro: Elsevier, 2018.

MIRZAEI, G. *et al*. An integrated data-mining and multi-criteria decision-making approach for hazard-based object ranking with a focus on landslides and floods. *Environmental Earth Sciences*, v. 77, n. 16, p. 1–23, 2018.

MONDAL, S.; MANDAL, S. Landslide susceptibility and risk: a micro level study from the Balason River basin in Darjeeling Himalaya. *Arabian Journal of Geosciences*, v. 11, n. 9, 2018.

MOORE, I. D.; GRAYSON, R. B.; LADSON, A. R. Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. *Hydrological Processes*, v. 5, n. 1, p. 3–30, 1991.

MORENO, C. C. Natural Language Processing, or How to Communicate With Your Computer. *Journal of the American College of Radiology*, p. 1–2, 2019.

MOUSAVI, S. Z. *et al*. GIS-based spatial prediction of landslide susceptibility using logistic regression model. *Geomatics, Natural Hazards and Risk*, v. 2, n. 1, p. 33–50, 2011.

MUÑOZ, P. *et al*. Flash-flood forecasting in an andean mountain catchment-development of a step-wise methodology based on the random forest algorithm. *Water (Switzerland)*, v. 10, n. 11, 2018.

NADKARNI, P. M.; OHNO-MACHADO, L.; CHAPMAN, W. W. Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, v. 18, n. 5, p. 544–551, 2011.

NANDI, A. *et al*. Flood hazard mapping in Jamaica using principal component analysis and logistic

regression. *Environmental Earth Sciences*, v. 75, n. 6, 2016.

NASCIMENTO, K. R. D. S.; ALENCAR, M. H. Management of risks in natural disasters: A systematic review of the literature on NATECH events. *Journal of Loss Prevention in the Process Industries*, v. 44, p. 347–359, 2016.

NE 10. **Chuvas provocam deslizamentos de barreira no Grande Recife**. Disponível em: <https://noticias.ne10.uol.com.br/grande-recife/noticia/2016/05/09/chuvas-provocam-deslizamentos-de-barreira-no-grande-recife-613637.php%0A>. Acesso em: 30 out. 2019.

NEFESLIOGLU, H. A. *et al*. Medium-scale hazard mapping for shallow landslide initiation: The Buyukkoy catchment area (Cayeli, Rize, Turkey). *Landslides*, v. 8, n. 4, p. 459–483, 2011.

NGUYEN, Q. K. *et al*. A novel hybrid approach based on instance based learning classifier and rotation Forest ensemble for spatial prediction of rainfall-induced shallow landslides using GIS. *Sustainability (Switzerland)*, v. 9, n. 5, 2017.

OLIVEIRA, P. T. S. *et al*. Curve number estimation from Brazilian Cerrado rainfall and runoff data. *Journal of Soil and Water Conservation*, v. 71, n. 5, p. 420–429, 2016.

OZDEMIR, A. Landslide susceptibility mapping using Bayesian approach in the Sultan Mountains (Akşehir, Turkey). *Natural Hazards*, v. 59, n. 3, p. 1573–1607, 2011.

P. DOMINGOS; PAZZANI, M. On the Optimality of the Simple Bayesian Classifier underZero-One Loss. Machine Learning. *Machine Learning*, v. 29, p. 103–130, 1997.

PAN, T. Y. *et al*. Hybrid neural networks in rainfall-inundation forecasting based on a synthetic potential inundation database. *Natural Hazards and Earth System Science*, v. 11, n. 3, p. 771–787, 2011.

PANWAR, V.; SEN, S. **Economic Impact of Natural Disasters: An Empirical Re-examination**. [s.l: s.n.]. v. 13

PAPATHOMA-KÖHLE, M. *et al*. Matrices, curves and indicators: A review of approaches to assess physical vulnerability to debris flows. *Earth-Science Reviews*, v. 171, n. November 2016, p. 272–288, 2017.

PARK, I.; LEE, S. Spatial prediction of landslide susceptibility using a decision tree approach: a case study of the Pyeongchang area, Korea. *International Journal of Remote Sensing*, v. 35, n. 16, p. 6089–6112, 2014.

PENG, L. *et al*. Landslide susceptibility mapping based on rough set theory and support vector machines: A case of the Three Gorges area, China. *Geomorphology*, v. 204, p. 287–301, 2014.

PERROCA, M. G.; GAIDZINSKI, R. R. Avaliando a confiabilidade interavaliadores de um instrumento para classificação de pacientes--coeficiente kappa. *Revista da Escola de Enfermagem da U S P*, v. 37, n. 1, p. 72–80, 2003.

PHAM, B. T. *et al*. A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India). *Environmental Modelling and Software*, v. 84, p. 240–250, 2016a.

PHAM, B. T. *et al*. Rotation forest fuzzy rule-based classifier ensemble for spatial prediction of landslides using GIS. *Natural Hazards*, v. 83, n. 1, p. 97–127, 2016b.

PHAM, B. T. *et al*. A novel ensemble classifier of rotation forest and Naïve Bayer for landslide susceptibility assessment at the Luc Yen district, Yen Bai Province (Viet Nam) using GIS. *Geomatics, Natural Hazards and Risk*, v. 8, n. 2, p. 649–671, 2017.

PHAM, B. T. A Novel Classifier Based on Composite Hyper-cubes on Iterated Random Projections for Assessment of Landslide Susceptibility. *Journal of the Geological Society of India*, v. 91, n. 3, p. 355–362, 2018.

PHAM, B. T.; PRAKASH, I.; TIEN BUI, D. Spatial prediction of landslides using a hybrid machine learning approach based on Random Subspace and Classification and Regression Trees. *Geomorphology*, v. 303, p. 256–270, 2018.

PHAM, B. T.; TIEN BUI, D.; PRAKASH, I. Landslide Susceptibility Assessment Using Bagging Ensemble Based Alternating Decision Trees, Logistic Regression and J48 Decision Trees Methods: A Comparative Study. *Geotechnical and Geological Engineering*, v. 35, n. 6, p. 2597–2611, 2017.

PHAM, B. T.; TIEN BUI, D.; PRAKASH, I. Bagging based Support Vector Machines for spatial prediction of landslides. *Environmental Earth Sciences*, v. 77, n. 4, p. 1–17, 2018.

POLYKRETIS, C.; CHALKIAS, C. Comparison and evaluation of landslide susceptibility maps obtained from weight of evidence, logistic regression, and artificial neural network models. *Natural Hazards*, v. 93, n. 1, p. 249–274, 2018.

POLYKRETIS, C.; CHALKIAS, C.; FERENTINOU, M. Adaptive neuro-fuzzy inference system (ANFIS) modeling for landslide susceptibility assessment in a Mediterranean hilly area. *Bulletin of Engineering Geology and the Environment*, v. 78, n. 2, p. 1173–1187, 2019.

POLYKRETIS, C.; FERENTINOU, M.; CHALKIAS, C. A comparative study of landslide susceptibility mapping using landslide susceptibility index and artificial neural networks in the Krios River and Krathis River catchments (northern Peloponnesus, Greece). *Bulletin of Engineering Geology and the Environment*, v. 74, n. 1, p. 27–45, 2014.

PORTAL FOLHAPE. **Deslizamento de barreira atinge quatro casas em Dois Unidos**. Disponível em: <https://www.folhape.com.br/noticias/noticias/cotidiano/2018/11/15/NWS,87656,70,449,NOTICIAS,2190-DESLIZAMENTO-BARREIRA-ATINGE-QUATRO-CASAS-DOIS-UNIDOS.aspx>. Acesso em: 28 out. 2019.

POUDYAL, C. P. *et al*. Landslide susceptibility maps comparing frequency ratio and artificial neural networks: A case study from the Nepal Himalaya. *Environmental Earth Sciences*, v. 61, n. 5, p. 1049–1064, 2010.

POURGHASEMI, H. R. *et al*. Assessment of landslide-prone areas and their zonation using logistic regression, LogitBoost, and naïvebayes machine-learning algorithms. *Sustainability (Switzerland)*, v. 10, n. 10, 2018.

POURGHASEMI, H. R.; KERLE, N. Random forests and evidential belief function-based landslide susceptibility assessment in Western Mazandaran Province, Iran. *Environmental*

Earth Sciences, v. 75, n. 3, p. 1–17, 2016.

POURGHASEMI, H. R.; MORADI, H. R.; FATEMI AGHDA, S. M. Landslide susceptibility mapping by binary logistic regression, analytical hierarchy process, and statistical index models and assessment of their performances. *Natural Hazards*, v. 69, n. 1, p. 749–779, 2013.

POURGHASEMI, H. R.; RAHMATI, O. Prediction of the landslide susceptibility: Which algorithm, which precision? *Catena*, v. 162, n. October 2017, p. 177–192, 2018.

POURGHASEMI, H. R.; ROSSI, M. Landslide susceptibility modeling in a landslide prone area in Mazandarn Province, north of Iran: a comparison between GLM, GAM, MARS, and M-AHP methods. *Theoretical and Applied Climatology*, v. 130, n. 1–2, p. 609–633, 2017.

PRABU, S.; RAMAKRISHNAN, S. S. Combined use of socio economic analysis, remote sensing and GIS data for landslide hazard mapping using ANN. *Journal of the Indian Society of Remote Sensing*, v. 37, n. 3, p. 409–421, 2009.

PRADHAN, B. Remote sensing and GIS-based landslide hazard analysis and cross-validation using multivariate logistic regression model on three test areas in Malaysia. *Advances in Space Research*, v. 45, n. 10, p. 1244–1256, 2010.

PRADHAN, B.; LEE, S. Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. *Environmental Modelling and Software*, v. 25, n. 6, p. 747–759, 2010.

PRADHAN, B.; PUTRA, U. LANDSLIDE SUSCEPTIBILITY MAPPING USING SUPPORT VECTOR MACHINE AND GIS AT THE GOLESTAN PROVINCE, IRAN. n. July 2016, p. 349–369, 2013.

PRANCKEVIČIUS, T.; MARCINKEVIČIUS, V. Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. *Baltic Journal of Modern Computing*, v. 5, n. 2, p. 221–232, 2017.

QGIS DEVELOPMENT TEAM. **QGIS Geographic Information System: Open Source Geospatial Foundation Project**, 2019. Disponível em: <http://qgis.osgeo.org>. Acesso em: 20 jun. 2019.

R CORE TEAM. **R: A language and environment for statistical computing**Vienna, AustriaR Foundation for Statistical Computing, , 2019. Disponível em: <https://www.r-project.org/> Acesso em: 25 mai. 2019.

RADAIDEH, Q. A. AL; KHATEEB, S. S. AL. An associative rule-based classifier for Arabic medical text. *International Journal of Knowledge Engineering and Data Mining*, v. 3, n. 3/4, p. 255, 2015.

RAHMATI, O.; POURGHASEMI, H. R. Identification of Critical Flood Prone Areas in Data-Scarce and Ungauged Regions: A Comparison of Three Data Mining Models. *Water Resources Management*, v. 31, n. 5, p. 1473–1487, 2017.

RAMAKRISHNAN, D. *et al*. Soft computing and GIS for landslide susceptibility assessment in Tawaghat area, Kumaon Himalaya, India. *Natural Hazards*, v. 65, n. 1, p. 315–330, 2013.

RAMANI, S. E.; PITCHAIMANI, K.; GNANAMANICKAM, V. R. GIS based landslide susceptibility mapping of Tevankarai Ar sub-watershed, Kodaikkanal, India using binary logistic regression analysis. *Journal of Mountain Science*, v. 8, n. 4, p. 505–517, 2011.

RAMANI SUJATHA, E.; KUMARAVEL, P.; RAJAMANICKAM G, V. Landslide Susceptibility Mapping Using Remotely Sensed Data through Conditional Probability Analysis Using Seed Cell and Point Sampling Techniques. *Journal of the Indian Society of Remote Sensing*, v. 40, n. 4, p. 669–678, 2012.

RAMASUBRAMANIAN, K.; SINGH, A. **Machine Learning Using R: With Time Series and Industry-Based Use Cases in R**. 2ª ed. New Delhi: Apress, 2017.

RAZAVI TERMEH, S. V. *et al*. Flood susceptibility mapping using novel ensembles of adaptive neuro fuzzy inference system and metaheuristic algorithms. *Science of the Total Environment*, v. 615, p. 438–451, 2018.

REGMI, A. D. *et al*. Landslide susceptibility mapping along Bhalubang — Shiwapur area of mid-Western Nepal using frequency ratio and conditional probability models. *Journal of Mountain Science*, v. 11, n. 5, p. 1266–1285, 2014.

ROBIN, X. *et al*. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, v. 12, p. 77, 2011.

ROSSI, M. *et al*. Optimal landslide susceptibility zonation based on multiple forecasts. *Geomorphology*, v. 114, n. 3, p. 129–142, 2010.

SABBAH, T. *et al*. Modified frequency-based term weighting schemes for text classification. *Applied Soft Computing Journal*, v. 58, p. 193–206, 2017.

SAMANTA, R. K. *et al*. Flood susceptibility mapping using geospatial frequency ratio technique: a case study of Subarnarekha River Basin, India. *Modeling Earth Systems and Environment*, v. 4, n. 1, p. 395–408, 2018.

SAMANTA, S.; PAL, D. K.; PALSAMANTA, B. Flood susceptibility analysis through remote sensing, GIS and frequency ratio model. *Applied Water Science*, v. 8, n. 2, p. 1–14, 2018.

SAMUEL, A. L. Some Studies in Machine Learning Using the game of Checkers. *IBM Journal*, v. 3, n. 3, p. 534–554, 1959.

SANGCHINI, E. K. *et al*. Assessment and comparison of combined bivariate and AHP models with logistic regression for landslide susceptibility mapping in the Chaharmahal-e-Bakhtiari Province, Iran. *Arabian Journal of Geosciences*, v. 9, n. 3, 2016.

SCIARRA, M.; COCO, L.; URBANO, T. Assessment and validation of GIS-based landslide susceptibility maps: a case study from Feltrino stream basin (Central Italy). *Bulletin of Engineering Geology and the Environment*, v. 76, n. 2, p. 437–456, 2017.

SECRETARIA EXECUTIVA DE DEFESA CILVIL. **Manual Técnico de Defesa Civil**. Recife, PE: Secretaria Executiva de Defesa Civil, 2012.

SEDEC. **Dados Abertos da Prefeitura de Recife**. Disponível em: <http://dados.recife.pe.gov.br>. Acesso em: 8 maio. 2019.

SEDEC - SECRETARIA EXECUTIVA DE DEFESA CIVIL. **SECRETARIA-EXECUTIVA DE DEFESA CIVIL**. Disponível em: <http://www2.recife.pe.gov.br/pagina/secretaria-executiva-de-defesa-civil>. Acesso em: 10 fev. 2019.

SEGONI, S. *et al*. Integration of rainfall thresholds and susceptibility maps in the Emilia Romagna (Italy) regional-scale landslide warning system. *Landslides*, v. 12, n. 4, p. 773–785, 2015.

SHAFIZADEH-MOGHADAM, H. *et al*. Novel forecasting approaches using combination of machine learning and statistical models for flood susceptibility mapping. *Journal of Environmental Management*, v. 217, p. 1–11, 2018.

SHAHABI, H.; HASHIM, M.; AHMAD, B. BIN. Remote sensing and GIS-based landslide susceptibility mapping using frequency ratio, logistic regression, and fuzzy logic methods at the central Zab basin, Iran. *Environmental Earth Sciences*, v. 73, n. 12, p. 8647–8668, 2015.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. **Understanding machine learning: From theory to algorithms**. [s.l: s.n.]. v. 9781107057

SHARMA, L. P. *et al*. Application of frequency ratio and likelihood ratio model for geo-spatial modelling of landslide hazard vulnerability assessment and zonation: a case study from the Sikkim Himalayas in India. *Geocarto International*, v. 29, n. 2, p. 128–146, 2014.

SHRESTHA, S.; KANG, T.-S.; SUWAL, M. An Ensemble Model for Co-Seismic Landslide Susceptibility Using GIS and Random Forest Method. *ISPRS International Journal of Geo-Information*, v. 6, n. 11, p. 365, 2017.

SILVA, E. L. DA; MENEZES, E. M. **Metodologia da Pesquisa e Elaboração de Dissertação**. 4. ed. Florianópolis: UFSC, 2005.

SMITH, L. *et al*. Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework. *Journal of Flood Risk Management*, v. 10, n. 3, p. 370–380, 2017.

SOLAIMANI, K.; MOUSAVI, S. Z.; KAVIAN, A. Landslide susceptibility mapping based on frequency ratio and logistic regression models. *Arabian Journal of Geosciences*, v. 6, n. 7, p. 2557–2569, 2013.

STEFANIDIS, S.; STATHIS, D. Assessment of flood hazard based on natural and anthropogenic factors using analytic hierarchy process (AHP). *Natural Hazards*, v. 68, n. 2, p. 569–585, 2013.

SU, C. *et al*. Mapping of rainfall-induced landslide susceptibility in Wencheng, China, using support vector machine. *Natural Hazards*, v. 76, n. 3, p. 1759–1779, 2015.

SUJATHA, E. R. *et al*. Landslide susceptibility analysis using probabilistic likelihood ratio model-a geospatial-based study. *Arabian Journal of Geosciences*, v. 6, n. 2, p. 429–440, 2013.

SUN, X. *et al*. Landslide Susceptibility Mapping Using Logistic Regression Analysis along the Jinsha River and Its Tributaries Close to Derong and Deqin County, Southwestern China. *ISPRS International Journal of Geo-Information*, v. 7, n. 11, p. 438, 2018.

TANER SAN, B. An evaluation of SVM using polygon-based random sampling inlandslide susceptibility mapping: The Candir catchment area(western Antalya, Turkey). *International*

Journal of Applied Earth Observation and Geoinformation, v. 26, n. 1, p. 399–412, 2014.

TANJIN AMIN, M.; KHAN, F.; AMYOTTE, P. A Bibliometric Review of Process Safety and Risk Analysis. *Process Safety and Environmental Protection*, 2019.

TEHRANY, M. S. *et al*. Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. *Catena*, v. 125, p. 91–101, 2015a.

TEHRANY, M. S. *et al*. Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. *Catena*, v. 125, p. 91–101, 2015b.

TEHRANY, M. S.; PRADHAN, B.; JEBUR, M. N. Flood susceptibility analysis and its verification using a novel ensemble support vector machine and frequency ratio method. *Stochastic Environmental Research and Risk Assessment*, v. 29, n. 4, p. 1149–1165, 2015.

TEN VELDHUIS, J. A. E.; HARDER, R. C.; LOOG, M. Automatic classification of municipal call data to support quantitative risk analysis of urban drainage systems. *Structure and Infrastructure Engineering*, v. 9, n. 2, p. 141–150, 2013.

TIEN BUI, D. *et al*. Landslide susceptibility assessment in vietnam using support vector machines, decision tree, and nave bayes models. *Mathematical Problems in Engineering*, v. 2012, 2012a.

TIEN BUI, D. *et al*. Spatial prediction of landslide hazards in Hoa Binh province (Vietnam): A comparative assessment of the efficacy of evidential belief functions and fuzzy logic models. *Catena*, v. 96, p. 28–40, 2012b.

TIEN BUI, D. *et al*. Landslide susceptibility assessment in the Hoa Binh province of Vietnam: A comparison of the Levenberg-Marquardt and Bayesian regularized neural networks. *Geomorphology*, v. 171–172, p. 12–29, 2012c.

TIEN BUI, D. *et al*. Hybrid artificial intelligence approach based on neural fuzzy inference model and metaheuristic optimization for flood susceptibilitgy modeling in a high-frequency tropical cyclone area using GIS. *Journal of Hydrology*, v. 540, p. 317–330, 2016a.

TIEN BUI, D. *et al*. GIS-based modeling of rainfall-induced landslides using data mining-based functional trees classifier with AdaBoost, Bagging, and MultiBoost ensemble frameworks. *Environmental Earth Sciences*, v. 75, n. 14, 2016b.

TIEN BUI, D. *et al*. Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides*, v. 13, n. 2, p. 361–378, 2016c.

TIEN BUI, D. *et al*. A novel fuzzy K-nearest neighbor inference model with differential evolution for spatial prediction of rainfall-induced shallow landslides in a tropical hilly area using GIS. *Landslides*, v. 14, n. 1, p. 1–17, 2017.

TRIGILA, A. *et al*. Comparison of Logistic Regression and Random Forests techniques for shallow landslide susceptibility assessment in Giampilieri (NE Sicily, Italy). *Geomorphology*, v. 249, p. 119–136, 2015.

TSAI, C. H.; CHEN, C. W. An earthquake disaster management mechanism based on risk assessment information for the tourism industry-a case study from the island of Taiwan.

Tourism Management, v. 31, n. 4, p. 470–481, 2010.

TSAI, F. *et al*. Analysis of topographic and vegetative factors with data mining for landslide verification. *Ecological Engineering*, v. 61, p. 669–677, 2013.

TSANGARATOS, P. *et al*. Applying Information Theory and GIS-based quantitative methods to produce landslide susceptibility maps in Nancheng County, China. *Landslides*, v. 14, n. 3, p. 1091–1111, 2017.

TSANGARATOS, P.; ILIA, I. Landslide susceptibility mapping using a modified decision tree classifier in the Xanthi Perfection, Greece. *Landslides*, v. 13, n. 2, p. 305–320, 2016.

TUKEY, J. K. **Exploratory Data Analysis**. Reading, Massachusetts: Addison-Wesley, 1977.

TURRIONI, J. B.; MELLO, C. H. P. **METODOLOGIA DE PESQUISA EM ENGENHARIA DE PRODUÇÃO: ESTRATÉGIAS, MÉTODOS E TÉCNICAS PARA CONDUÇÃO DE PESQUISAS QUANTITATIVAS E QUALITATIVAS**. Itajubá: UNIFEI, 2012.

TV OUL. **Chuva provoca deslizamento de barreira na Zona Norte do Recife**. Disponível em: <https://tvuol.uol.com.br/video/chuva-provoca-deslizamento-de-barreira-na-zona-norte-do-recife-0402CC983164E0996326%0A>. Acesso em: 30 out. 2019.

TYAGI, J. V. *et al*. SCS-CN based time-distributed sediment yield model. *Journal of Hydrology*, v. 352, n. 3–4, p. 388–403, 2008.

UNISDR. **Living with Risk: A global review of disaster reduction initiatives**. New York and Geneva: [s.n.]. v. 2

UNITED STATES GEOLOGICAL SURVEY - USGS. **Landsat 8**. Disponível em: <https://landsat.gsfc.nasa.gov/landsat-8/>. Acesso em: 20 out. 2019.

VALERIANO, M. D. M. **Topodata: Guia Para Utilização De Dados**. São José dos Campos, SP: Instituto Nacional de Pesquisas Espaciais (INPE), 2008.

WANG, L. J. *et al*. Landslide susceptibility mapping in Mizunami City, Japan: A comparison between logistic regression, bivariate statistical analysis and multivariate adaptive regression spline models. *Catena*, v. 135, p. 271–282, 2015a.

WANG, L. J.; SAWADA, K.; MORIGUCHI, S. Landslide susceptibility analysis with logistic regression model based on FCM sampling strategy. *Computers and Geosciences*, v. 57, p. 81–92, 2013.

WANG, P. *et al*. GIS-based random forestweight for rainfall-induced landslide susceptibility assessment at a humid region in Southern China. *Water (Switzerland)*, v. 10, n. 8, 2018.

WANG, Q. *et al*. GIS based frequency ratio and index of entropy models to landslide susceptibility mapping (Daguan, China). *Environmental Earth Sciences*, v. 75, n. 9, 2016.

WANG, Q. *et al*. Integration of information theory, K-Means cluster analysis and the logistic regression model for landslide susceptibility mapping in the three gorges area, China. *Remote Sensing*, v. 9, n. 9, 2017.

WANG, Z. *et al*. Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, v. 527, p. 1130–1141, 2015b.

WANG, Z. *et al*. Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, v. 527, p. 1130–1141, 2015c.

WICKHAM, H. **Tidyverse: Easily Install and Load the "Tidyverse"**, 2017. Disponível em: <https://cran.r-project.org/package=tidyverse> Acesso em: 15 mai. 2019.

WIPO. **WIPO Technology Trends 2019: Artificial Intelligence**. Genebra, Suiça: World Intellectual Property Organization, 2019.

XIA, H. *et al*. Subpixel inundation mapping using landsat-8 OLI and UAV data for a wetland region on the zoige plateau, China. *Remote Sensing*, v. 9, n. 1, p. 1–22, 2017.

XIE, Z. *et al*. A comparative study of landslide susceptibility mapping using weight of evidence, logistic regression and support vector machine and evaluated by SBAS-InSAR monitoring: Zhouqu to Wudu segment in Bailong River Basin, China. *Environmental Earth Sciences*, v. 76, n. 8, p. 1–19, 2017.

XU, C. *et al*. Application of an incomplete landslide inventory, logistic regression model and its validation for landslide susceptibility mapping related to the May 12, 2008 Wenchuan earthquake of China. *Natural Hazards*, v. 68, n. 2, p. 883–900, 2013.

XU, J. *et al*. Natural disasters and social conflict: A systematic literature review. *International Journal of Disaster Risk Reduction*, v. 17, p. 38–48, 2016.

YAO, X.; THAM, L. G.; DAI, F. C. Landslide susceptibility mapping based on Support Vector Machine: A case study on natural slopes of Hong Kong, China. *Geomorphology*, v. 101, n. 4, p. 572–582, 2008.

YEON, Y. K.; HAN, J. G.; RYU, K. H. Landslide susceptibility mapping in Injae, Korea, using a decision tree. *Engineering Geology*, v. 116, n. 3–4, p. 274–283, 2010.

YI-TING, W. *et al*. USING STATISTICAL LEARNING ALGORITHMS IN REGIONAL LANDSLIDE SUSCEPTIBILITY ZONATION WITH LIMITED LANDSLIDE FIELD DATA. *Journal of M Mountain Scien e*, v. 12, n. 2, p. 268–288, 2015.

YILMAZ, I. Landslide susceptibility mapping using frequency ratio, logistic regression, artificial neural networks and their comparison: A case study from Kat landslides (Tokat-Turkey). *Computers and Geosciences*, v. 35, n. 6, p. 1125–1138, 2009a.

YILMAZ, I. A case study from Koyulhisar (Sivas-Turkey) for landslide susceptibility mapping by artificial neural networks. *Bulletin of Engineering Geology and the Environment*, v. 68, n. 3, p. 297–306, 2009b.

YILMAZ, I. Comparison of landslide susceptibility mapping methodologies for Koyulhisar, Turkey: Conditional probability, logistic regression, artificial neural networks, and support vector machine. *Environmental Earth Sciences*, v. 61, n. 4, p. 821–836, 2010a.

YILMAZ, I. The effect of the sampling strategies on the landslide susceptibility mapping by conditional probability and artificial neural networks. *Environmental Earth Sciences*, v. 60, n.

3, p. 505–519, 2010b.

YOUSSEF, A. M. *et al*. Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. *Landslides*, v. 13, n. 5, p. 839–856, 2016.

YOUSSEF, A. M.; PRADHAN, B.; HASSAN, A. M. Flash flood risk estimation along the St. Katherine road, southern Sinai, Egypt using GIS based morphometry and satellite imagery. *Environmental Earth Sciences*, v. 62, n. 3, p. 611–623, 2011.

ZARE, M. *et al*. Landslide susceptibility mapping at Vaz Watershed (Iran) using an artificial neural network model: A comparison between multilayer perceptron (MLP) and radial basic function (RBF) algorithms. *Arabian Journal of Geosciences*, v. 6, n. 8, p. 2873–2888, 2013.

ZHANG, K. *et al*. The assessment of landslide susceptibility mapping using random forest and decision tree methods in the Three Gorges Reservoir area, China. *Environmental Earth Sciences*, v. 76, n. 11, 2017a.

ZHANG, K. *et al*. The assessment of landslide susceptibility mapping using random forest and decision tree methods in the Three Gorges Reservoir area, China. *Environmental Earth Sciences*, v. 76, n. 11, p. 1–20, 2017b.

ZHAO, G. *et al*. Mapping flood susceptibility in mountainous areas on a national scale in China. *Science of the Total Environment*, v. 615, p. 1133–1142, 2018.

ZHU, A. X. *et al*. Comparison of the presence-only method and presence-absence method in landslide susceptibility mapping. *Catena*, v. 171, n. July, p. 222–233, 2018.

APÊNDICE A – PROTOCOLO DA REVISÃO

Tema	**Mapeamento de perigo de Inundação e deslizamento de terra com métodos de aprendizado de máquina**
Autor	Saulo Guilherme Rodrigues
Aprovação	Marcelo Hazin Alencar

1. Justificativa para a pesquisa

A presente revisão sistemática da literatura tem como objetivo principal identificar as principais características do mapeamento de perigo de inundação e deslizamento utilizando métodos de aprendizado de máquina. Dentre os resultados esperados, encontram-se:

1. Responder as questões de pesquisa estabelecidas para essa pesquisa;
2. Formalizar um framework para mapeamento de perigo de inundações e deslizamentos baseado em métodos de aprendizado de máquina.

Além disso, foi evidenciado através de uma busca prévia na literatura que não há, até então, uma revisão sistemática da literatura publicada com as características da proposta do presente trabalho.

2. Questões de pesquisa

A presente pesquisa busca responder as seguintes questões de pesquisa:

- **RQ1:** Quais métodos de aprendizado de máquina são mais utilizados para o mapeamento do perigo de enchentes e inundações?
- **RQ2:** Como estão distribuídos os valores da avaliação de performance dos métodos de aprendizado de máquina para o mapeamento do perigo de enchentes e inundações?
- **RQ3:** Quais as variáveis condicionantes mais utilizadas nos métodos de aprendizado de máquina para o mapeamento do perigo de enchentes e inundações?
- **RQ4:** Como estão distribuídos os tamanhos das amostras utilizadas nos métodos de aprendizado de máquina para o mapeamento do perigo de enchentes e inundações?
- **RQ5:** Como se deu a validação do modelo de aprendizado de máquina para o mapeamento do perigo de enchentes e inundações?

3. Estratégia de busca

A busca por estudos preliminares será realizada na base de dados Web of Science Core Collection. As buscas serão realizadas nos campos título, resumo, palavras chaves e palavras chaves mais.

Os termos de busca foram divididos em dois grupos. O primeiro referente as técnicas de aprendizado de máquina, os quais foram tirados de literatura específica. Já o segundo, referente a inundações e deslizamentos. Os termos utilizados estão representados no Quadro A.1.

Quadro A.1 – Termos de busca

Grupo relacionado a machine learning	Grupos relacionados aos desastres
"Machine Learning" OR "Deep-Learning" OR "data mining" OR "artificial intelligence" OR "nearest neighbo*" OR "K-NN" OR "decision tree*" OR " linear regression" OR "regression tree*" OR "classification trees" OR "neural network*" OR "ANN" OR "genetic algorithm" OR "association rule*" OR "support vector machine*" OR "SVM" OR "support vector regression" OR "random forest" OR "boosting"; OR "ensemble learning" OR "ensemble model*" OR "gradient descent" OR "clustering" OR "logistic regression" OR "genetic algorithm" OR "naive bayes" OR "bagging" OR "Least-square support vector machines" OR "K-Means" OR "Dimensionality Reduction" OR "boosting" OR "adaboost" OR "Principal component analysis" OR "Classifcation and Regression Tree" OR "classifcation rules" OR "Association Rules" OR "Linear Discriminant Analysis"	*"Flood* prediction" OR "Flood* vulnerability" OR "Flood* estimation" OR "Flood* forecast" OR "Flood* analysis" OR "Flood* susceptibility" OR "Flood* assessment"OR "Flood* hazard" OR "Flood* risk" OR* "Landslid * *vulnerability" OR* "Landslid* *prediction" OR* "Landslid* *estimation" OR* "Landslid* *forecast" OR* "Landslid* *analysis" OR "*Landslid* *susceptibility" OR "*Landslid* *assessment"OR* "Landslid* *hazard" OR* "Landslid* *risk" OR* "inundation * *prediction" OR* "inundation *estimation" OR* "inundation *forecast" OR* "inundation *analysis" OR* "inundation * *vulnerability" OR "*inundation *susceptibility" OR "*inundation *assessment" OR* "inundation *hazard" OR* "inundation *risk"*

4. Critério de inclusão

Quadro A.2 – Critérios de inclusão

Grupo	Critérios
Características do documento	• Somente Artigos publicados em periódicos; • Somente artigos escritos em língua inglesa; • Somente artigos publicados até 31/12/2018.
Área do estudo	Serão consideradas somente as seguintes áreas de estudo: • Geologia; • Recurso hídricos; • Engenharia; • Geologia; • Ciências ambientais e ecologia; • Ciências meteorológicas; • Geografia física; • Sensoriamento remoto; • Ciência da computação; • Matemática; • Tecnologia outros tópicos; • Pesquisa operacional e ciência da gestão.
Tipo do estudo	• Artigos que mapeiam o perigo de enchentes e deslizamentos de terra utilizando métodos de aprendizado de máquina; • Somente artigos que realizam o mapeamento espacial do perigo.

5. Procedimentos para seleção dos estudos

Os estudos serão selecionados com base nos critérios de inclusão e com o escopo previamente definido. Todos os trabalhos serão selecionados pelo pesquisador com a supervisão do orientador da pesquisa.

6. Estratégia da extração dos dados

Os dados serão extraídos com o auxílio do pacote desenvolvido em R Bibliometrix (ARIA; CUCCURULLO, 2017). Tal ferramenta fornece um banco de dados com 37 variáveis diferentes, contendo informações sobre cada trabalho incluído na pesquisa, dentre elas:

- **AU**: Autores;
- **TI**: Título;
- **SO**: Fonte;
- **JI**: Abreviação ISSO para a fonte;
- **DT**: Tipo do documento;
- **DE**: Palavras chaves;
- **AB**: Resumo;
- **C1**: Nacionalidade do autor;
- **CR**: Referências citadas;
- **TC**: Número de citações;
- **PY**: Ano;
- **SC**: Categoria.

Além das análises proporcionadas por tal ferramenta, serão extraídas as seguintes informações complementares, totalizando 42 informações analisadas:

- Método de aprendizado de máquina utilizado com maior performance;
- Valor da performance;
- Variáveis condicionantes adotadas no modelo;
- Tamanho da amostra utilizada;
- Maneira de validação do modelo.

7. Síntese dos dados extraídos

A análise dos dados extraídos será realizada com o auxílio de pacotes elaborados na linguagem R, são eles:

- ***Tidyverse***: Pacote que fornece uma vasta gama de ferramentas para a importação, limpeza, tratamento e visualização de dados;
- ***Bibliometrix***: pacote fornece ferramentas específicas para a análise bibliométrica, incluindo análise gráficas e estatísticas.

8. Estratégia de disseminação.

A disseminação dos resultados da pesquisa será realizada através da comunicação cientifica. Publicações em congressos e periódicos em forma de artigos. Publicação em forma de dissertação de mestrado no Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Pernambuco, campus acadêmico do agreste.