Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Ciência da Computação

Predição de Escorregamentos de Encostas baseada em Aprendizado de Máquina

Laedson Silva Pedreira

Feira de Santana

2022

Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Ciência da Computação

Laedson Silva Pedreira

Predição de Escorregamentos de Encostas baseada em Aprendizado de Máquina

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Orientador: Dr. Rodrigo Tripodi Calumby

Coorientador: Dr.ª Maria do Socorro Costa São Mateus

Feira de Santana

2022

Ficha catalográfica - Biblioteca Central Julieta Carteado - UEFS

Pedreira, Laedson Silva
P399p Predição de escorregamentos de encostas baseada em aprendizado de máquina / Laedson Silva Pedreira. - 2022.
124f. : Il.

Orientador: Rodrigo Tripodi Calumby

Dissertação (mestrado) - Universidade Estadual de Feira de Santana. Programa de Pós-Graduação em Ciência da Computação, 2022.

1. Encostas - Escorregamentos. 2. Deslizamento de terra. 3. Aprendizado de máquina. 4. Mineração de dados (computação). 5. LightGBM. 6. Random Forest. I. Calumby, Rodrigo Tripodi, orient. II. Universidade Estadual de Feira de Santana. III. Título.

CDU: 004.85:624.137.2

Rejane Maria Rosa Ribeiro – Bibliotecária CRB-5/695

Laedson Silva Pedreira

Predição de Escorregamentos de Encostas baseada em Aprendizado de Máquina

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Feira de Santana, 16 de março de 2022

BANCA EXAMINADORA

Rodrigo Tripodi Calumby (Orientador(a))
Universidade Estadual de Feira de Santana

Carlos César Uchôa de Lima
Universidade Estadual de Feira de Santana

Angelo Conrado Loula
Universidade Estadual de Feira de Santana

Abstract

Landslides are among the main phenomena that cause natural disasters across the planet. Every year landslides have caused numerous material damages and claimed a large number of fatalities. In order to understand and describe the phenomenon of landslides, in addition to preventing or minimizing the problems caused by them, many studies have been carried out on their dynamics. However, considering the complexity of the problem and the scarcity of integrated and large-scale data, specific studies of individualized predictive models and with a temporal relationship, for monitoring and indicating risks are challenging. Despite this, the application of predictive models based on machine learning has great potential to contribute with effective and efficient tools, capable of assisting in the monitoring and prevention of damages arising from such events. In this context, this work proposes and experimentally evaluates data mining and machine learning techniques for the construction of a database from multiple sources, its pre-processing and the prediction of landslides individually, in time and in space. In addition, in order to verify the impact on the predictive capacity of the classifiers, the implications of two methods of generating non-slip samples, the number of days of accumulated rainfall considered and the lead time of prediction were analyzed. With the application of the methodology proposed here, it was possible to predict landslides in a promising way, with *F1-score* values greater than 0.929 ± 0.002 and AUC greater than 0.930 ± 0.002. The results presented also suggest that the use of these predictive models can contribute to a better decision-making by the competent about the regarding the monitoring and prevention of damage caused by landslides induced by rain.

Keywords: Landslide, Random Forest, Machine Learning, Data Mining, Prediction, LightGBM.

Resumo

Os escorregamentos de encostas constituem um dos principais fenômenos causadores de desastres naturais em todo planeta. Todos os anos os escorregamentos têm causado inúmeros prejuízos materiais e fazendo um grande número de vítimas fatais. Com o intuito de compreender e descrever o fenômeno dos escorregamentos, além de prevenir ou minimizar os problemas por eles causados, muitos estudos têm sido realizados acerca da sua dinâmica. Contudo, considerando-se a complexidade do problema e escassez de dados integrados e em larga escala, estudos específicos de modelos preditivos individualizados e com relação temporal, para monitoramento e indicação de riscos são desafiadores. Apesar disso, a aplicação de modelos preditivos baseados em aprendizado de máquina apresenta grande potencial em contribuir com ferramentas eficazes e eficientes, capazes de auxiliar no monitoramento e prevenção de danos oriundos de tais eventos. Neste contexto, este trabalho propõe e avalia experimentalmente técnicas de mineração de dados e aprendizado de máquina para a construção de uma base de dados a partir de múltiplas fontes, seu pré-processamento e a predição de escorregamentos de encostas de forma individualizada, no tempo e no espaço. Além disso, a fim de verificar o impacto na capacidade preditiva dos classificadores, foram analisadas as implicações de dois métodos de geração de amostras de não escorregamentos, do número de dias de chuva acumulada considerada e do tempo de antecedência de predição. Com a aplicação da metodologia aqui proposta foi possível realizar predição de escorregamentos de modo promissor, com valores de *F1-score* superiores a $0,929 \pm 0,002$ e AUC superiores a $0,930 \pm 0,002$. Os resultados apresentados sugerem ainda que a utilização desses modelos preditivos pode contribuir para uma melhor tomada de decisão dos órgãos competentes no que se refere ao monitoramento e prevenção de danos causados pelos escorregamentos de encostas induzidos por chuva.

Palavras-chave: Escorregamento, *Random Forest*, Aprendizado de Máquina, Mineração de dados, Predição, *LightGBM*, Deslizamento de terra.

Prefácio

Esta dissertação de mestrado foi submetida à Universidade Estadual de Feira de Santana (UEFS) como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

A dissertação foi desenvolvida no Programa de Pós-Graduação em Ciência da Computação (PGCC), tendo como orientador o Prof. Dr. **Dr. Rodrigo Tripodi Calumby**. O Prof. Dr. **Dr.ª Maria do Socorro Costa São Mateus** foi coorientador(a) deste trabalho.

Agradecimentos

Primeiramente, agradeço a **Deus**, porque é a luz, fortaleza, proteção e sabedoria que dá sentido à minha vida. Obrigado, Senhor, por me dar saúde para a batalha do dia-a-dia, sem Ti nada posso fazer.

Aos meus pais, **José Carneiro** e **Marizete**, que nunca mediram esforços para me ensinar o caminho do bem, e sempre me apoiaram em todas as etapas da minha vida. Sem vocês, eu não chegaria até aqui. Muito obrigado por tudo! O amor que sinto por vocês é incondicional.

À minha amada esposa **Nathália Karolin** e as minhas filhas **Maria Clara** e **Ana Júlia**, por todo amor, incentivo, apoio e compreensão. Nada disso teria sentido se vocês não existissem na minha vida.

Ao ilustre **Prof. Dr. Rodrigo Tripodi Calumby**, pela orientação, competência, profissionalismo e dedicação. Tenho certeza que não chegaria a este ponto sem o seu apoio. Meu muito obrigado e que Deus o abençoe grandemente.

À ilustre **Prof.ª Dr.ª Maria do Socorro Costa São Mateus** por toda ajuda durante a realização deste trabalho. Muito obrigado!

Aos membros da banca examinadora, **Prof. Dr. Angelo Conrado Loula** e **Prof. Dr. Carlos César Uchoa de Lima**, que tão gentilmente aceitaram participar e colaborar com esta dissertação.

A todos os **professores do PPCC** que foram de suma importância para a transmissão do conhecimento através de debates, críticas e outros métodos.

Aos **colegas de mestrado** que me acompanharam durante todo este percurso, proporcionando uma valiosa troca de experiências e conhecimentos que hoje faz parte de minha vida.

À **Defesa Civil de Salvador** (CODESAL), ao **Laboratório de Geotecnia da UFBA** e a **Companhia de Pesquisa de Recursos Minerais** (CPRM) pela atenção e pela disponibilização dos dados utilizados no trabalho.

A todos que de forma direta ou indireta contribuíram para que a conclusão deste Mestrado se tornasse possível.

"O que fazemos em vida ecoa na eternidade."

– Autor desconhecido

Sumário

Alinhamento com a Linha de Pesquisa

Linha de Pesquisa: Computação Inteligente

Na dissertação são aplicados conceitos, técnicas e ferramentas da Inteligência Artificial com o objetivo de auxiliar o entendimento e a resolução da problemática dos deslizamentos de terra. São tratados temas como Aprendizagem de Máquina e Mineração de Dados.

Produções Bibliográficas, Produções Técnicas e Premiações

1. PEDREIRA, L. S.; CALUMBY, R. T. ; MATEUS, M. S. C. S. . Predição de Escorregamentos de Encostas baseada em Aprendizado de Máquina. In: PESQBASE - Workshop de Pesquisa da Escola Regional de Computação Bahia-Alagoas-Sergipe, 2021, Maceió. Anais do PESQBASE - Workshop de Pesquisa da Escola Regional de Computação Bahia-Alagoas-Sergipe.
2. PEDREIRA, L. S.; MATEUS, M. S. C. S. ; CALUMBY, R. T. . Integração de sistemas para predição de deslizamentos de terra baseada em aprendizado de máquina. In: SBSI - Simpósio Brasileiro de Sistemas de Informação, 2022, Curitiba. Anais SBSI - Simpósio Brasileiro de Sistemas de Informação. Curitiba, 2022.

Lista de Tabelas

Lista de Figuras

Lista de Abreviações

Abreviação	**Descrição**
MDE	Modelo Digital de Elevação
SIG	Sistema de Informações Geográficas
LGBM	*Light Gradient Boosting Machine*
RNA	Redes Neurais Artificiais
AD	Árvore de Decisão
RF	*Random Forest*
SVM	Support Vector Machines
MLP	*Multi-Layer Perceptron*
RBF	*Radial Basis Function*
KLR	*Kernel Logistic Regression*
LMT	*Logistic Model Trees*
SC	Single CART
NBT	*Naive Bayes Trees*
LR	*Logistic Regression*
RSS	*Random Subspace*
CART	*Classification And Regression Trees*
BRT	*Boosted Regression Tree*
GBRT	*Gradient Boosted Regression Tree*
ROC	*Receiver Operating Characteristic*
AUC	Área sob a Curva ROC
ACC	Acurácia
KDD	*Knowledge Discovery in Databases*
VP	Verdadeiro Positivo
FN	Falso Negativo
FP	Falso Positivo
VN	Verdadeiro Negativo
TVP	Taxa de Verdadeiros Positivos
TFP	Taxa de Falsos Positivos
GOSS	Gradient-based One-Side Sampling
EFB	Exclusive Feature Bundling
WSR	Wilcoxon signed-ranks
ONU	Organização das Nações Unidas

IBGE	Instituto Brasileiro de Geografia e Estatística
CPRM	Companhia de Pesquisa de Recursos Minerais
CEMADEC	Centro de Monitoramento e Alerta da Defesa Civil de Salvador
CODESAL	Defesa Civil de Salvador
CEMADEN	Centro Nacional de Monitoramento e Alertas de Desastres Naturais
INMET	Instituto Nacional de Meteorologia
INEMA	Instituto do Meio Ambiente e Recursos Hídricos
EM-DAT	Banco de Dados Internacional de Desastres

Capítulo 1

Introdução

"Comece fazendo o que é necessário, depois o que é possível, e de repente estará fazendo o impossível."

– São Francisco de Assis

Os movimentos de massa constituem um relevante processo natural que atuam na dinâmica das encostas, fazendo parte da evolução geomorfológica em regiões de relevo acidentado. O crescimento desordenado do espaço urbano em áreas desfavoráveis, sem o adequado planejamento do uso do solo e sem a adoção de técnicas de engenharia para promover a estabilização, está disseminando a ocorrência de acidentes associados a esses processos, que muitas vezes atingem dimensões de desastres (Tominaga et al., 2009).

Os escorregamentos de encostas, um dos tipos de movimentos de massa, também conhecidos como deslizamentos de terra, ocorrem por todo planeta e vêm provocando, todos os anos, inúmeros problemas para a sociedade, tais como a destruição de edificações e equipamentos urbanos, prejuízos em atividades produtivas, impactos ambientais e fazendo um grande número de vítimas fatais (IBGE, 2019). De acordo com o Banco de Dados Internacional de Desastres (EM-DAT), do Centro de Pesquisa em Epidemiologia de Desastres, os escorregamentos são responsáveis por cerca de 17% das fatalidades associadas aos desastres naturais todos os anos (Aristizábal e Sánchez, 2020), e, segundo a ONU (1993), os eventos de movimentos de massa está entre os fenômenos naturais que mais causam prejuízos financeiros e mortes no mundo.

Os escorregamentos também constituem um dos principais fenômenos causadores de desastres naturais nas cidades brasileiras (Tominaga et al., 2009), e são potencializados, no caso do Brasil, principalmente pela ação das águas das chuvas, que

infiltram no solo acima das encostas, provocando a redução da sua estabilidade e, consequentemente, induzindo os escorregamentos (Farah, 2003).

Entre os anos de 2007 e 2017, foram registradas no país 1.756 mortes em virtude de eventos de deslizamentos de terra (Macedo, 2019). Vários episódios podem ilustrar os problemas gerados pelos deslizamentos no Brasil, bem como mostram a influência da precipitação como fator desencadeante desses eventos. Devido às fortes chuvas, em 2008, a cidade de Blumenau, em Santa Catarina, chegou a registrar mais de três mil deslizamentos de terra, com 24 mortes e deixando 5.209 pessoas desabrigadas. Em todo o Estado, 9.390 pessoas ficaram desalojadas e houve 135 mortes (Oliveira et al., 2018). No *réveillon* de 2009, fortes chuvas provocaram 31 deslizamentos no município de Angra dos Reis, no Estado do Rio de Janeiro, sendo registradas 53 mortes (Lauriano, 2010). Em 2011, enchentes e deslizamentos de terra atingiram o Estado do Rio de Janeiro, deixando 918 mortos e, pelo menos, 99 pessoas desaparecidas (Knust, 2021). Em 2015, devido às precipitações pluviométricas, um escorregamento de encosta deixou cerca de 502 pessoas desabrigadas e 15 vítimas fatais na cidade de Salvador, Bahia (Amado, 2015). Também, devido às fortes chuvas do mês de janeiro de 2022, o Estado de São Paulo registrou 34 mortes provocadas por deslizamentos de terra (Henrique, 2022).

O município de Salvador, no Estado da Bahia, registra expressivas ocorrências de deslizamentos que afetam centenas de pessoas todos os anos. Como em outras cidades brasileiras, a problemática das encostas de Salvador se dá, também, devido à ocupação desordenada de suas áreas com relevo acidentado, por meio de edificações construídas sem adoção dos critérios técnicos necessários. Segundo a Defesa Civil (CODESAL), a cidade tem 400 áreas de risco e, nestes locais, são mais de 1.000 pontos de perigo, o que evidencia um problema significativo para a cidade (Souza, 2015).

Impedir que os deslizamentos de terra ocorram, segundo Kobiyama et al. (2006), foge da capacidade humana, sendo necessária a adoção de medidas de prevenção que minimizem os impactos causados por eles. Contudo, medidas estruturais de prevenção envolvem obras de engenharia que, em geral, são complexas e de alto custo. Outras medidas podem ser implementadas e geralmente envolvem ações de planejamento e gerenciamento, como zoneamento ambiental e sistemas de alerta. O primeiro é difícil de ser implementado na prática, em virtude da dificuldade do poder público em fiscalizar e controlar o crescimento desordenado de ocupações em áreas de risco. No caso da existência de atividades humanas já implantadas em áreas suscetíveis aos deslizamentos de terra, a criação de um sistema de alerta nestes espaços pode auxiliar na redução dos danos materiais e na preservação de vidas.

Desta forma, a previsão de deslizamentos de terra pode ser extremamente útil para minimizar o número de vítimas e reduzir os danos materiais, de forma a dar suporte ao poder público no processo de tomada de decisões e no gerenciamento situacional, a fim de reduzir os danos causados por esses fenômenos. Contudo, o desenvolvimento de uma abordagem para previsão de escorregamento de encostas é muito desafia-

dora. Primeiramente devido à baixa disponibilidade de dados integrados e em larga escala, além de que uma estimativa precisa, envolve um número considerável de variáveis inter-relacionadas, o que torna a previsão de escorregamentos de encostas um problema não trivial e não linear (Tehrani et al., 2019).

Neste contexto, alguns modelos matemáticos foram desenvolvidos com o intuito de prever a suscetibilidade de escorregamentos, como, por exemplo, os modelos SHALSTAB (Montgomery e Dietrich, 1994), SINAP (Pack et al., 1998), TRIGRS (Baum et al., 2008), entre outros, que utilizam uma base física para produzir mapas a partir de equações que buscam simular os mecanismos deflagradores dos escorregamentos. Esses modelos são baseados na teoria de talude infinito, elaborada por Mohr-Coulomb, que define as tensões responsáveis pela desestabilização de uma parte do solo na encosta (Carson et al., 1972). Desta forma, esses modelos buscam apontar, preliminarmente, as possíveis áreas susceptíveis aos escorregamentos, não sendo capazes de prever a ocorrência futura de um evento.

De acordo com Tominaga et al. (2009), a associação dos escorregamentos à estação das chuvas, notadamente às chuvas intensas, já é de conhecimento generalizado. Com isso, alguns trabalhos importantes buscaram estabelecer uma relação numérica entre a precipitação e os escorregamentos (Elbachá et al. 1992; Castro 2006a; Parizzi et al. 2010; D'Orsi 2011), entre outros. Esses trabalhos procuram, por meio da Estatística, determinar uma relação entre a intensidade e a duração das chuvas com o fenômeno de escorregamentos de encostas. Contudo, esses modelos não levam em consideração as propriedades geológicas e geomorfométricas das encostas e não têm a intenção de prever o escorregamento de forma específica no tempo e no espaço, mas estabelecer limiares de chuvas que desencadeiam os eventos de deslizamentos de terra.

Embora a mecânica básica possa ser utilizada em modelos físicos e numéricos para estabelecer a relação entre a resistência ao cisalhamento do solo e as tensões cisalhantes mobilizadoras que determinam a estabilidade, ou não, de uma encosta, a escassez de medições consideravelmente detalhadas e em tempo real das condições do solo, do maciço rochoso e das águas subterrâneas, impede a previsão precisa dos escorregamentos. Desta forma, os pesquisadores estão explorando cada vez mais técnicas relacionadas à utilização de mineração de dados e aprendizado de máquina na tentativa de aproximar a ocorrência de escorregamentos futuros a padrões de distribuição anteriores, em função do grande potencial dessas técnicas em descobrir relações complexas entre os dados com múltiplos componentes associados (Korup e Stolle, 2014).

Entretanto, a maior parte desses trabalhos compreende a confecção de mapas de susceptibilidade, mas não a previsão futura desses eventos. São escassos os estudos que buscam prever escorregamentos de encostas de modo individualizado e com relação temporal (Souza e Ebecken, 2012; Farahmand e Aghakouchak, 2013; Tehrani et al., 2019), devido à baixa disponibilidade de dados precisos de deslizamentos (geolocalização e data), fatores desencadeantes (precipitação pluviométricas) e fatores de controle (geomorfologia, geologia, solo etc). Com isso, é notório que essa área de

aplicação ainda se encontra em desenvolvimento, com vários desafios e limitações a serem superadas.

Nesse sentido, em função da complexidade do contexto apresentado e dos desafios computacionais a serem superados, a construção de um modelo preditivo de escorregamento de encostas passa pela obtenção de dados, garantia e melhoria de qualidade dos dados, integração dos dados, escolha de algoritmos, construção, otimização e validação de modelos eficazes e avaliação de cenários de aplicação.

Com isso, considerando a análise da literatura e o contexto de aplicação, diversas perguntas ainda precisam ser respondidas, de forma que este estudo se propõe explicar as questões enumeradas a seguir:

1. Quais os dados disponíveis e quais podem ser úteis de forma a integrá-los e prepará-los para uso em aprendizado de máquina?
2. Qual a capacidade dos algoritmos em aprender com estes dados e qual sua eficácia preditiva?
3. Como gerar amostras de não escorregamentos e qual o impacto de diferentes métodos na capacidade preditiva dos algoritmos?
4. Qual a implicação na capacidade preditiva dos modelos ao se considerar diferentes espaços temporais de antecedência de previsão?
5. Qual a quantidade de dias de chuva acumulada é mais efetiva na predição de escorregamentos?
6. A utilização de métodos de classificadores *ensemble*[1] aumenta a performance de modelos preditivos de escorregamento de encosta quando comparados com classificadores tradicionais, visto o potencial desses métodos em conferir melhores resultados preditivos?

Assim, este estudo propõe a aplicação de técnicas de mineração de dados e aprendizado de máquina na construção de uma base de dados integrada e na predição de escorregamento de encostas induzidos por chuvas, buscando responder algumas lacunas de conhecimento na área de aplicação, por meio da utilização de classificadores com diferentes métodos de construção, bem como avaliar diferentes metodologias de geração de amostras de não ocorrências, diferentes intervalos de tempo para a chuva acumulada anterior ao escorregamento e simular diferentes intervalos de antecedência preditiva.

1.1 Objetivos

Este trabalho tem como objetivo propor e avaliar a aplicação técnica de aprendizado de máquina e mineração de dados, no pré-processamento e na predição de

[1] Método de aprendizado de máquina que combina o resultado de múltiplos modelos em busca de produzir um melhor resultado preditivo (Kelleher et al., 2015).

escorregamento de encostas induzidos por chuvas, integrando os dados das ocorrências passadas com as propriedades do solo, dados geomorfométricos e a precipitação pluviométrica. Para isso, serão comparados múltiplos algoritmos de aprendizado de máquina com diferentes características e avaliar diferentes cenários de aplicação. De maneira a alcançar o objetivo geral deste estudo, será necessário:

1. Aplicar técnicas de mineração de dados para extrair e analisar informações de variados conjuntos de dados;
2. Construir um *dataset* com as características essenciais para o desenvolvimento do modelo de predição de escorregamento de encostas;
3. Avaliar e aplicar técnicas de pré-processamento, para tratamento e transformação de dados;
4. Realizar otimização dos modelos preditivos;
5. Comparar os modelos preditivos de classificação, com o intuito de verificar qual denota as melhores taxas de acerto, considerando múltiplos cenários de estudo e modelagem do problema.
6. Avaliar o impacto dos métodos de geração de amostras de não deslizamentos na capacidade preditivas dos modelos preditivos.

1.2 Contribuições

A previsão dos eventos de deslizamento de terra é de grande importância para a mobilização das instituições responsáveis em assistir a população que ocupa as encostas. O desenvolvimento dos modelos propostos, utilizando técnicas de mineração de dados e aprendizado de máquina, deverá utilizar, de forma automatizada, inúmeros dados relativos ao fenômeno dos deslizamentos, gerando informações cruciais para tomada de decisão.

Além disso, destaca-se o potencial desta pesquisa em contribuir diretamente com a construção de base de dados integrada, a avaliação de eficácia de modelos preditivos, a verificação de diferentes espaços temporais de antecedência de previsão e a análise de pontos de falhas e sugestões de melhorias na coleta de dados. Por fim, pode-se enumerar algumas contribuições indiretas ou posteriores ao desenvolvimento deste trabalho:

1. Aumento da assertividade na estimativa de escorregamentos de encostas;
2. Melhoria no monitoramento das áreas de riscos de escorregamentos;
3. Incremento na qualidade das informações úteis para o processo de tomada de decisão;
4. Simulação de risco de escorregamentos conforme previsão de chuvas;
5. Redução das perdas humanas e materiais.

1.3 Organização do Trabalho

As demais partes deste documento estão organizadas como é detalhado a seguir:

O capítulo 2 apresenta o conhecimento que fundamenta este estudo de forma a embasar a metodologia e os resultados. Veremos neste capítulo a teoria relacionada aos fenômenos dos escorregamentos de encostas, continuando com os estudos da mineração de dados e do aprendizado de máquina e, por fim, será discutido como essas técnicas têm sido aplicadas nos problemas relacionados à predição de escorregamentos de terra;

O capítulo 3 detalha a metodologia do trabalho, sendo descritas as características dos dados, as metodologias para geração dos cenários de análise, as configurações utilizadas na geração dos modelos e como foram avaliados experimentalmente.

No capítulo 4, serão apresentados os resultados obtidos pelo estudo e aplicação da metodologia descrita anteriormente. Também serão analisados e discutidos os resultados alcançados, com o objetivo de avaliar e inferir conclusões e limitações acerca dos modelos estudados.

O Capítulo 5 expõe as conclusões de todo o estudo, além de possíveis trabalhos futuros.

Capítulo 2

Revisão Bibliográfica

"O que sabemos é uma gota; o que ignoramos é um oceano."
– Isaac Newton

Neste capítulo serão apresentados os elementos fundamentais dos eventos de escorregamentos de encostas, bem como a definição, os mecanismos deflagadores e as variáveis mais importantes a serem analisadas. Ademais, serão apresentados os conceitos e técnicas de mineração de dados e aprendizado de máquina empregados neste estudo. Ao final do capítulo, serão apresentados alguns trabalhos com o objeto de estudo relacionado ao desta dissertação.

2.1 Movimentos de Massa

Wicander e Monroe (2009) definem movimentos de massa como o deslocamento, encosta abaixo, de solo e fragmentos de rochas devido à ação da gravidade. Eles estabelecem que fatores como a declividade da encosta, intemperismo, clima, conteúdo da água, vegetação e sobrecarga são inter-relacionados e afetam a movimentação de massas.

A implacável e ininterrupta gravidade é a principal força por trás da movimentação de massas. Desta forma, o colapso de uma encosta ocorre quando a força da gravidade que age sobre ela supera a força de coesão da camada de sedimentos resultantes da fragmentação de rochas subjacentes ou da própria rocha (resistência à deformação). Os fatores que ajudam a manter a estabilidade são a declividade e a coesão do material da encosta, o atrito entre os grãos e qualquer sustentação externa da encosta. Esses fatores, coletivamente, definem a resistência ao cisalhamento da encosta (Wicander e Monroe, 2009).

De acordo com Augusto Filho e Virgili (1998), a definição do tipo de movimento de massa está diretamente ligada às características e condições locais onde são consideradas, como por exemplo, a estrutura geológica, a declividade do terreno, a forma e orientação da encosta, a área de contribuição, a distribuição e a intensidade das chuvas.

A diversidade de materiais, processos e fatores condicionantes e desencadeadores que envolvem os movimentos de massas, levaram a um grande interesse por sistemas classificação. As propostas mais adotadas para descrição e classificação dos movimentos de massa incluem os trabalhos de Terzaghi (1950), Varnes (1978) e Augusto Filho (1992). A classificação de Varnes (1978) se tornou referência para muitos pesquisadores ao redor do mundo e se baseou no tipo de movimento e material transportado, identificando seis categorias principais de movimentos de massa que são apresentadas na Tabela 2.1.

Tabela 2.1: Classificação dos movimentos de massa segundo (Varnes, 1978)

<table>
<tr><th rowspan="3">Tipo de Movimento</th><th colspan="3">Tipo de Material</th></tr>
<tr><th rowspan="2">Rocha</th><th colspan="2">Solo em Engenharia</th></tr>
<tr><th>Predominantemente Grosseiro</th><th>Predominantemente Fino</th></tr>
<tr><td>Queda</td><td>de rocha</td><td>de detritos</td><td>de solo</td></tr>
<tr><td>Tombamentos</td><td>de rocha</td><td>de detritos</td><td>de solo</td></tr>
<tr><td>Escorregamentos</td><td>de rocha</td><td>de detritos</td><td>de solo</td></tr>
<tr><td>Expansão Lateral</td><td>de rocha</td><td>de detritos</td><td>de solo</td></tr>
<tr><td rowspan="2">Escoamentos (*)</td><td rowspan="2">de rocha
(rastejo profundo)</td><td>de detritos</td><td>de solo</td></tr>
<tr><td colspan="2">Rastejo de solo</td></tr>
<tr><td>Complexos</td><td colspan="3">Combinação de dois ou mais tipos de movimentos, ação de vários agentes (simultâneos ou sucessivos)</td></tr>
</table>

(*) o autor subdivide os diferentes tipos de corridas de acordo com a velocidade e conteúdo de água dos materiais.

No Brasil, pode-se destacar as classificações desenvolvidas por Freire (1965) e Guidicini e Nieble (1984). Também deve ser mencionada a classificação proposta por Augusto Filho (1992), que, assim como a do IPT (1991), define os tipos de movimentos de massa como sendo: rastejos, escorregamentos, quedas e corridas. Neste trabalho será utilizada a classificação proposta por Augusto Filho (1992) (Tabela 2.2).

Esta pesquisa abordará apenas os movimentos de massa do tipo escorregamentos, visto que são os mais frequentes no Brasil, para os quais foram obtidas as informações necessárias para o desenvolvimento dos modelos preditivos.

No Brasil, os escorregamentos em encostas naturais e artificiais constituem os principais fenômenos causadores de desastres naturais. Todos os anos esses eventos provocam inúmeras perdas econômicas e impactos ambientais, além de gerar um número significativo de mortes. Por esse motivo, a ocorrência dos referidos escorregamentos em áreas urbanas ou mesmo em áreas desabitadas, precisa ser monitorada de forma preventiva, justificando a necessidade de ferramentas eficazes e eficientes que deem suporte ao poder público na gestão dos riscos.

Tabela 2.2: Características dos principais movimentos de massa (Augusto Filho, 1992).

Processos	Características, Material e Geometria
Rastejos	Vários planos de deslocamento (internos) Velocidades muito baixas (cm/ano) a baixas e decrescentes com a profundidade Movimentos constantes, sazonais ou intermitentes Solo, depósitos, rocha alterada/fraturada Geometria indefinida
Escorregamentos	Poucos planos de deslocamento (externos) Velocidades médias (m/h) a altas (m/s) Pequenos a grandes volumes de material Geometria e materiais variáveis Planares – solos pouco espessos, solos e rochas com um plano de fraqueza Circulares – solos espessos homogêneos e rochas muito fraturadas Em cunha – solos e rochas com dois planos de fraqueza
Quedas	Sem planos de deslocamento Movimentos tipo queda livre ou em plano inclinado Velocidades muito altas (vários m/s) Material rochoso Pequenos a médios volumes Geometria variável: lascas, placas, blocos, etc. Rolamento de matacão e tombamento.
Corridas	Muitas superfícies de deslocamento Movimento semelhante ao de um líquido viscoso Desenvolvimento ao longo das drenagens Velocidades médias a altas Mobilização de solo, rocha, detritos e água Grandes volumes de material Extenso raio de alcance, mesmo em áreas planas

2.2 Escorregamentos

De acordo com Tominaga et al. (2009), os escorregamentos de terra, consistem em movimentos rápidos, de porções de terreno (solos e rochas), com volumes definidos, deslocando-se sob ação da gravidade, para baixo e para fora do talude ou da encosta. Segundo Farah (2003), eles são potencializados principalmente pela ação das águas de chuvas, que infiltram e saturam o solo das encostas, resultando na redução da sua estabilidade.

Dentre os tipos de movimento de massa, os escorregamentos têm recebido uma atenção particular da comunidade científica nos últimos anos, levando em consideração os enormes problemas causados e agravados pela ocupação desordenada das áreas de encostas (Guimarães et al., 2003).

De acordo com Tominaga et al. (2009), quando se leva em consideração a geometria e a natureza dos materiais instabilizados, os escorregamentos podem ser subdivididos em três tipos: escorregamentos rotacionais ou circulares, escorregamentos translacionais ou planares e escorregamentos em cunha. Abaixo serão descritas as principais

características dos tipos de escorregamentos.

2.2.1 Escorregamentos rotacionais ou circulares

Conforme Tominaga et al. (2009), os escorregamentos rotacionais ou circulares são caracterizados por uma superfície de ruptura circular, com concavidade voltada para cima, ao longo do qual ocorre o movimento rotacional do maciço de solo (Figura 2.1).

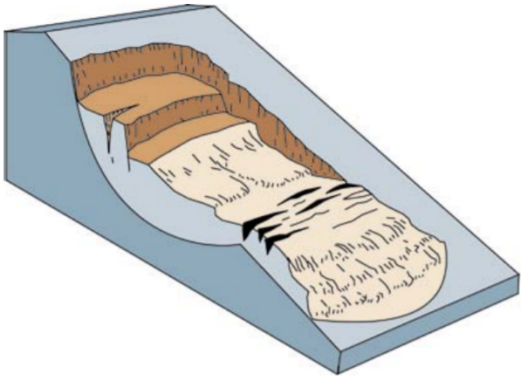

Figura 2.1: Escorregamento rotacional (Highland e Bobrowsky, 2008).

Os escorregamentos rotacionas estão associados, com maior frequência, a materiais homogêneos e solos espessos, como os decorrentes da alteração de rochas argilosas. Estão associados a taludes que variam de 20 a 40 graus em inclinação (Highland e Bobrowsky, 2008).

Diversas causas podem estar relacionadas a esse tipo de escorregamento, desde a interferência humana até aspectos geológicos e principalmente a ação da água, seja por meio da redução da sucção do solo ou através do surgimento de pressões positivas de água (Jesus, 2008).

2.2.2 Escorregamentos translacionais ou planares

De acordo com Tominaga et al. (2009), esses são os tipos de escorregamentos mais frequentes entre todos os tipos de movimentos de massa. Eles formam uma superfície de ruptura planar relacionadas às heterogeneidades dos solos e rochas que representam descontinuidades mecânicas e/ou hidrológicas oriundas de processos geológico, geomorfológico ou pedológicos (Figura 2.2).

Segundo Guidicini e Nieble (1984), são escorregamentos caracterizados por serem rasos com um plano de fraqueza, na maioria dos casos entre 0,5 a 5,0m de profundidade, com maiores extensões no comprimento. Ocorrem em encostas tanto de alta como de baixa declividade e o movimento é predominantemente acompanhado por uma translação.

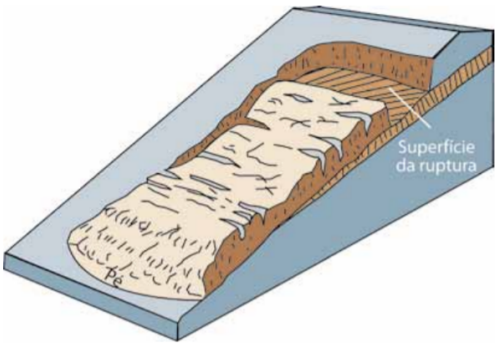

Figura 2.2: Escorregamento translacionais (Highland e Bobrowsky, 2008).

Longos períodos de chuva intensas, em geral, provocam os escorregamentos translacionais. Normalmente, a superfície de ruptura coincide com a interface solo-rocha, a qual representa uma relevante descontinuidade mecânica e hidrológica. A ação das águas nestes movimentos é mais superficial. As rupturas ocorrem em curto espaço de tempo devido ao rápido aumento da umidade durante os eventos pluviométricos de grande intensidade (Fernandes e Amaral, 1996 apud Tominaga et al., 2009).

De acordo com Wolle (1988), este tipo de escorregamento é frequente no sudeste brasileiro, especialmente na região da Serra do Mar , que se estende desde o estado do Rio de Janeiro até o estado do Rio Grande do Sul, onde as cicatrizes apresentadas após o escorregamento apresentam largura na faixa de 10 a 20 metros, com espessuras inferiores a 4 metros e comprimentos que podem atingir 20 metros.

2.2.3 Escorregamentos em cunha

Os escorregamentos do tipo cunha (Figura 2.3) têm sua ocorrência mais restrita à regiões em que o relevo é estritamente controlado por estruturas geológicas. Esse tipo de escorregamento está geralmente relacionado aos maciços rochosos, pouco ou muito alterados, nos quais a existência de duas estruturas planares, desfavoráveis à estabilidade, condiciona o deslocamento de um prisma ao longo do eixo de intersecção destes planos. Eles são mais comuns em taludes de corte ou em encostas que sofreram algum tipo de desconfiamento, natural ou antrópico (Tominaga et al., 2009).

2.3 Fatores condicionantes dos escorregamentos

Diversos fatores associados à geologia, geomorfologia, hidrogeologia da área e à ação antrópica, influenciam no processo que desestabiliza e desencadeia os escorregamentos. Desse modo, algumas causas influenciam mais diretamente do que outras, sendo de grande importância o seu conhecimento, já que possibilitam um melhor entendimento dos escorregamentos, como prevê-los, evitá-los ou se prevenir deles. Diversos autores discutem as relações entre esses fatores, dentre os quais podem-se destacar os

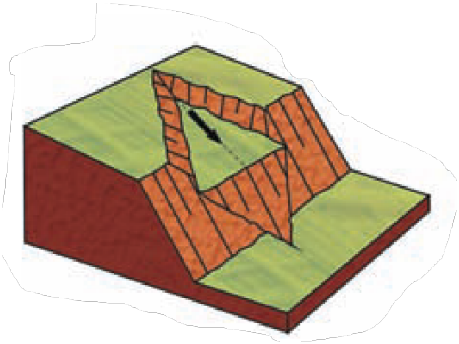

Figura 2.3: Escorregamentos em cunha (Tominaga et al., 2009).

trabalhos de Terzaghi (1950), Varnes (1978), Cruden e Varnes (1996), Augusto Filho e Virgili (1998).

Segundo Terzaghi (1950), as causas dos movimentos de massa podem ser subdivididas em duas categorias relacionadas aos taludes: causas internas e causas externas. As causas externas são aquelas que contribuem para o aumento das tensões cisalhantes devido às modificações da geometria do talude, entre outras. Enquanto que as causas internas são aquelas que provocam a diminuição da resistência ao cisalhamento do solo, como, por exemplo, a perda da resistência pela ação do intemperismo e a elevação da poropressão [1] na superfície potencial de escorregamento.

Varnes (1978), aborda o tema dividindo os principais condicionantes e mecanismos de deflagração dos escorregamentos em dois grupos: os fatores que aumentam as solicitações e os que diminuem a resistência do terreno da encostas, associado aos fenômenos naturais e antrópicos.

Cruden e Varnes (1996) sintetizaram os processos e as características que favorecem os movimentos de massa. Essas causas são apresentadas na Tabela 2.3, e estão separadas em quatro grupos práticos, conforme os procedimentos e ferramentas essenciais para que se possa iniciar uma investigação.

Augusto Filho e Virgili (1998), apresentaram os principais agentes condicionantes dos escorregamentos e processos relacionados aos eventos brasileiros, sendo eles:

- Características climatológicas, destacando-se o regime pluviométrico;
- Características e distribuição dos materiais que constituem o substrato das encostas/taludes;
- Características geomorfológicas, com destaque para inclinação e forma do perfil dos taludes;
- Regime das águas de superfície e subsuperfície;

[1] Pressão que um fluído exerce no interior dos poros dos elementos porosos como solos e rochas (de Sousa Pinto, 2016).

Tabela 2.3: Relação das causas de movimentos de massa Cruden e Varnes (1996).

Causas geológicas	**Causas físicas**
- Materiais fracos	- Chuvas intensas
- Materiais sensíveis	- Derretimento rápido de neve
- Materiais intemperizados	- Precipitações excepcionalmente prolongadas
- Materiais fissurados ou fraturados	- Terremotos
- Orientação desfavorável de descontinuidades (acamamento, xistosidade, etc.)	- Erupções vulcânicas
- Orientação desfavorável de descontinuidades estruturais (falhas, contatos, inconformidades, etc.)	- Descongelamento
- Contraste de permeabilidade	- Intemperismo por congelamento e descongelamento
- Contraste de rigidez (materiais densos, rígidos sobre materiais plásticos)	- Intemperismo por expansão e retração
Causas morfológicas	**Causas humanas**
- Levantamento tectônico ou vulcânico	- Escavações de taludes
- Alívio por degelo	- Sobrecarga no talude ou na crista
- Erosão fluvial no pé do talude	Rebaixamento (reservatórios)
- Erosão glacial no pé do talude	- Irrigação
- Erosão nas margens laterais	- Mineração
- Erosão subterrânea (Solução e *piping*)	- Vibração artificial
- Deposição de cargas no talude ou na crista	- Vazamento de água
- Remoção da vegetação (fogo, seca)	

- Características do uso e ocupação, incluindo cobertura vegetal e as diferentes formas de intervenção antrópicas das encostas, como cortes, aterros, concentração de água pluvial e servida, etc.

No caso dos deslizamentos de terra, é fundamental considerar conjuntamente aos condicionantes relacionados à geologia, à geomorfologia, às precipitações pluviométricas, e aos condicionantes relativos às ações antrópicas, para a ocorrência do evento de deslizamento. A ação antrópica é um fator agravante que contribui com profundas modificações no meio ambiente. Segundo Wolle (1988), as atividades antrópicas de uso e ocupação do solo constituem um agente que modifica a declividade, acelerando os processos que induzem os escorregamentos. Entre as atividades antrópicas modificadoras, podemos destacar o corte de taludes ou aterros e o desmatamento das encostas.

Nas próximas seções serão apresentados os condicionantes relacionados à geologia e geomorfologia e às precipitações pluviométricas.

2.3.1 Precipitações Pluviométricas

Em hidrologia, o termo precipitação é dado para toda água oriunda do meio atmosférico que atinge a superfície terrestre, tais como: neblina, granizo, geada, neve, orvalho e chuva. Devido à sua capacidade de gerar escoamento superficial, a chuva é o tipo de precipitação mais relevante para hidrologia (Tucci, 2020). Para o escopo deste trabalho, sempre que for utilizado o termo precipitação, este deve ser considerado equivalente à chuva.

De acordo com Tucci (2020), as principais características da precipitação são: o seu total, duração e distribuições temporal e espacial. Para ter significado, o total precipitado deve estar ligado a uma duração. Por exemplo, 100 mm pode ser pouco para um mês, mas é muito para um dia, ou, ainda mais, para uma hora. A ocorrência da precipitação é um processo aleatório que não permite uma previsão determinística com grande antecedência.

No que tange aos escorregamentos, Augusto Filho e Virgili (1998), afirmam que as chuvas relacionam-se diretamente com a dinâmica das águas de superfície e subsuperfície, influenciando a deflagração dos processos de instabilização de taludes e encostas. Para os autores, os índices pluviométricos críticos para a deflagração dos escorregamentos variam conforme o regime de infiltração no solo, a dinâmica das águas subterrâneas no solo e o tipo de instabilização.

Tatizana et al. (1987), ressaltam que os desastres naturais como escorregamentos, desabamentos e inundações são mais prováveis de ocorrerem durante os períodos de chuvas, visto que diminui a resistência do solo. Tatizana et al. (1987), afirma ainda que os limites de precipitação causadores de escorregamentos variam conforme as características de cada região analisada.

Highland e Bobrowsky (2008), estabelecem a saturação de água em encostas como a principal causa dos escorregamentos destas vertentes. Para eles, a saturação pode ocorrer sob diversas formas, sendo no Brasil a mais frequente pela ação das águas de chuvas. De acordo com Massad (2010), a infiltração das águas de chuva eleva as pressões neutras, o que reduz a resistência do solo, ou pode provocar a diminuição dos parâmetros de resistência, principalmente da coesão aparente.

A precipitação pluviométrica é, sem dúvida, um relevante fator condicionante dos escorregamentos. Em algumas regiões brasileiras, a associação dos escorregamentos às chuvas já é de conhecimento generalizado. Com isso, muitos autores tentaram estabelecer uma correlação entre as chuvas e os escorregamentos. Internacionalmente citam-se Endo (1969), Campbell (1975), Lumb (1975), Govi (1977), Eyles (1979), Kay e Chen (1995), Zêzere et al. (2003). No Brasil, diversos trabalhos tentaram estabelecer as variadas correlações entre a precipitação pluviométrica e os escorregamentos. Podem-se citar os estudos de Guidicini e Iwasa (1977), Tatizana et al. (1987) e Almeida et al. (1991) para Petrópolis, Elbachá et al. (1992) para Salvador, Xavier (1996) e Parizzi et al. (2010) para Belo Horizonte, Salaroli (2003) para Vitória, Vieira (2004) para Blumenau, Ide (2005) para Campinas, Castro (2006b)

para Ouro Preto, Salles e Amaral (2013) para a Região Serrana do Rio de Janeiro e Soares et al. (2015) para João Pessoa. As tabelas 2.5 e 2.4, citadas por Ide (2005), exibem de forma resumida algumas pesquisas e as suas respectivas conclusões.

Tabela 2.4: Resumo de pesquisas internacionais realizadas sobre chuva e escorregamento Ide (2005).

Autor e ano	**Local do estudo**	**Característica associada ao escorregamento**
Endo (1969)	Hokkaido	Limite de 200 mm / dia.
Campbell (1975)	Los Angeles	Limite de 262 mm / evento de chuva.
	Alameda County, Califórnia	Limite de 180 mm / evento de chuva.
Govi (1977)	Bacino Padano, Itália	Limite de 100 mm / 3 dias.
Eyles (1979)	Wellington City	Limite de 50-90 mm / evento de chuva.Escorregamentos de grande porte com 100 mm /evento de chuva.
Brand et al. (1984)	Hong Kong	Limite de 100 mm / 24 horas; 70 mm / hora.
Kay e Chen (1995)	Hong Kong	Relação: d = (180 – h)/ onde d é a chuva diária (mm), h é a chuva horária (mm) e s é o coeficiente de inclinação da reta que limita as zonas de probabilidade de ocorrência de escorregamentos.
Finlay et al. (1997)	Hong Kong	Boa relação com chuva de 1 e 12 horas anteriores. Limite de 8 a 17 mm / hora.
Zêzere et al. (2003)	Lisboa, Portugal	Limite de 220 mm / 15 dias para escorregamentos de pequeno porte / translacionais rasos. Limite de 130 mm / dia para escorregamentos múltiplos translacionais. Limites de 459 mm / 40 dias a 690 mm / 75 dias(chuvas prolongadas) para movimento de massa profundos.

Tabela 2.5: Resumos de pesquisas brasileiras realizadas sobre chuva e escorregamento. Tabela estendida pelo autor com base em Ide (2005).

Autor e ano	**Local do estudo**	**Característica associada a escorregamento**
Guidicini e Iwasa (1977)	Costa Ocidental, Brasil	Limite de 8 a 17% de pluviosidade anual. Com 20% da pluviosidade anual, desenvolvem-se fenômenos catastróficos.
Tatizana et al. (1987)	Serra do Mar, Brasil	Boa relação com precipitação acumulada de 4 dias;I (Ac) = 2603 Ac-0,933I = intensidade da precipitação Ac = precipitação acumulada de 4 dias
Elbachá et al. (1992)	Salvador, Bahia	Limite indicativo de 120 mm/ 4 dias.
Xavier (1996)	Belo Horizonte, Minas Gerias	Limite de 30 mm/24 horas e 50 mm em 48 horas.
GEO-RIO (2000)	Rio de Janeiro, Rio de Janeiro	Boa relação com chuva acumulada de 4 dias.
Salaroli (2003)	Vitória, Espirito Santo	36,00 mm para nível de Atenção. 87,5 mm para nível de Alerta
Vieira (2004)	Blumenau, Santa Catarina	Boa relação com chuva acumulada de 3 a 4 dias, somando em torno de 50 mm.
Ide (2005)	Campinas, São Paulo	78,0 mm para 7 dias de acumulada.
Castro (2006a)	Ouro Preto, Minas Gerais	129,0mm de chuva acumulada em cinco dias com chuvas diárias de 55,0mm.
Parizzi et al. (2010)	Belo Horizonte, Minas Gerais	3 dias iguais ou superiores a 100mm e chuvas diárias e intensas superiores a 70mm.;
Salles e Amaral (2013)	Região Serrana do Rio de Janeiro	Chuvas horárias muito fortes, acima de 55mm/h, ou de chuvas diárias acima de 120mm/24h. Chuvas da ordem de 30mm/h, 100mm/24h, 115mm/96h e 270mm/mês; C Região Serrana do Rio de Janeiro chuvas da ordem de 50mm/h, 120mm/24h, 130mm/96h e 300mm/mês.
Soares et al. (2015)	João Pessoa, Paraíba	50,0 mm de chuva acumulada de sete dias com chuvas diárias de 150,0 mm. 150,0 mm de chuva acumulada de sete dias com chuvas diárias de 50,0 mm.

A bibliografia apresentada nas tabelas 2.4 e 2.5, mostra que não existe um intervalo de tempo e um limiar ideal que possa caracterizar a relação entre a precipitação pluviométrica e os escorregamentos, tendo cada região seus limiares e intervalos de tempo mais prováveis. Isso ocorre devido às características geológicas e geomorfológicas de cada região que também favorecem a ocorrência dos fenômenos estudados.

2.3.2 Fatores Geomorfológicos

De acordo com Christofoletti (1980), a Geomorfologia é a ciência que estuda as formas do relevo. As formas determinam as particularidades espaciais de uma superfície, compondo as diferentes configurações de paisagem morfológica. É o seu

aspecto visível, a sua configuração que caracteriza o modelado topográfico de uma determinada área.

Sabe-se que os escorregamentos são condicionados por complexas relações entres diversos fatores. Para Fernandes et al. (2001), os fatores de natureza geomorfológica, denominado muitas vezes de parâmetros topográficos, tratam da relação entre a forma e a hidrologia (superficial e sub-superficial) da encosta, compreendendo parâmetros, tais como: declividade, forma da encosta, área de contribuição, orientação da encosta (aspecto), espessura do solo, comprimento da encosta, simetrias entre vales e a elevação.

Dentre os parâmetros citados, a declividade tem sido considerada como principal indicador de caráter topográfico nos estudos de previsão e definição de áreas instáveis (Fernandes et al., 2001).

Segundo Gramani e Kanji (2001), as encostas que apresentam elevadas declividades são as mais propícias à ocorrência de escorregamentos. Os autores consideram valores de aproximadamente 30° a 45° de declividade como valores críticos para estas ocorrências, embora isto seja variável para cada solo que compõe a encosta. Contudo, segundo os autores, são relatadas na bibliografia ocorrências de deslizamentos em encostas com inclinações em declividades baixas. Esse fato vai evidenciar que os fatores condicionantes estão inter-relacionados e que outros fatores também devem ser considerados.

2.3.3 Fatores Geológicos

De acordo com Augusto Filho e Virgili (1998), algumas características relacionadas ao ambiente geológico favorecem os escorregamentos. Essas características estão ligadas à litologia, padrões de fraturas e diaclases, às propriedades internas (textura e mineralogia), à coesão e o ângulo de atrito, à permeabilidade e ao manto de intemperismo.

De acordo com Brito (2014), é importante observar os aspectos da litologia na avaliação dos movimentos de massa, visto que cada litotipo apresenta diferentes graus de coesão, resistência e permeabilidade, o que irá influenciar o tipo de drenagem, a textura e a resistência da rocha aos processos de intemperismo. Do mesmo modo, a litologia influencia o tipo de solo que será gerado e, portanto, nas suas características geotécnicas.

Diversas propriedades dos solos influenciam o desencadeamento dos movimentos de massa, tais como: seu peso específico, porosidade, índices de vazios, mineralogia, granulometria, permeabilidade, compressibilidade, textura, coesão, ângulo de atrito, espessura, condutividade, hidráulica, histórico de tensões, entre outras. Essas propriedades estão diretamente ligadas à origem e formação do solo (Augusto Filho e Virgili, 1998). Segundo Guidicini e Nieble (1984), as propriedades mais significativas, quando se trata de movimentos de massa e estabilidade de taludes, são o ângulo de atrito e a coesão.

Os solos residuais podem apresentar várias descontinuidades, às quais se incluem feições estruturais reliquiares do embasamento rochoso (fraturas, falhas, bandamentos etc) e horizontes de solo formados por processos pedogenéticos. Essas descontinuidades facilitam a infiltração da água, provocando o aumento da pressão neutra no interior da encosta, o que leva à redução dos valores de resistência, e consequentemente, a sua desestabilização (Fernandes e Amaral, 2000).

Os depósitos de encostas, tanto na forma de tálus como de colúvios, estão diretamente associados à geomorfologia da encosta, essa combinação faz com que os depósitos de encostas assumam grande importância como condicionantes dos movimentos de massa. Esses materiais possuem grande heterogeneidade em função da descontinuidade espacial e temporal dos processos formadores desses depósitos. Muitos destes repousam diretamente na rocha sã, gerando descontinuidade mecânica e hidrológica. Ao longo desse contato, condições críticas de pressão neutra positivas podem ser alcançadas durante eventos pluviométricos, favorecendo a ocorrência de escorregamentos. (Fernandes e Amaral, 2000).

2.4 Descoberta de Conhecimento em Bancos de Dados

Vivemos em um mundo onde imensas quantidades de dados são produzidas e armazenadas diariamente. Analisar esses dados é uma necessidade importante e, para isso, ferramentas poderosas e versáteis que possam, por meio de técnicas e modelos computacionais, analisar essa grande quantidade de dados, são extremamente fundamentais para descobrir informações valiosas e transformar esses dados em conhecimento organizado. Essa necessidade levou ao surgimento da mineração de dados (Han et al., 2011), processo através do qual é possível descobrir padrões de conhecimento de interesse em grandes volumes de dados.

Este processo, também conhecido por *Knowledge Discovery in Databases* - (KDD), envolve a transformação de dados brutos em informações úteis, sendo constituído de várias etapas. Não é trivial e deve ser interativo, com o objetivo de identificar padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados (Tan et al. 2014; Fayyad et al. 1996).

De modo geral, a complexidade do processo de KDD está na dificuldade em entender e interpretar de maneira adequada os inúmeros fatos observáveis, bem como na dificuldade em conjugar dinamicamente tais interpretações, de forma a decidir quais ações que devem ser realizadas em cada situação. A difícil tarefa de guiar a execução do processo de KDD (Goldschmidt e Passos, 2005) fica ao encargo do analista humano .

Segundo Tan et al. (2014), o KDD é definido como um processo que compõe diversas etapas de transformações. A Figura 2.4 apresenta de maneira resumida as etapas realizadas em um processo KDD. A primeira etapa é o pré-processamento que tem como objetivo a transformação dos dados brutos de entrada em um formato apropriado para análise, por meio da seleção, limpeza, codificação e enriquecimento destes

dados. Na etapa de mineração de dados é executada, de fato, uma busca por padrões consistentes e/ou relacionamentos sistemáticos entre variáveis. Já a etapa de pós-processamento garante que apenas resultados válidos e úteis sejam incorporados no sistema de suporte à decisão.

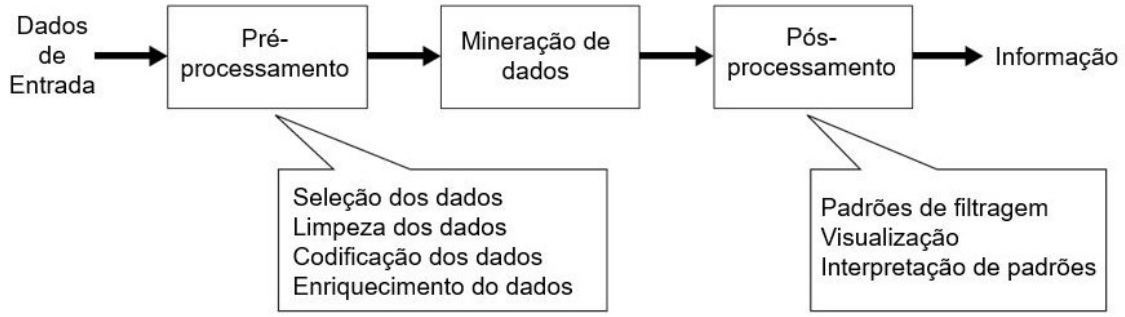

Figura 2.4: Etapas do processo KDD (Tan et al., 2014).

Para alguns autores, o KDD refere-se ao processo geral de descoberta de conhecimento útil a partir de dados e mineração de dados se refere a uma etapa específica nesse processo (Fayyad et al., 1996). Contudo, em algumas bibliografias, o termo mineração de dados (*Data Mining*) tornou-se mais popular que o KDD e é utilizado quando se refere ao processo de identificações de padrões a partir de grandes volumes de dados armazenados em banco de dados (Han et al., 2011).

2.4.1 Etapas do processo KDD

Pré-processamento

Segundo Tan et al. (2014), a etapa de pré-processamento consiste em transformar os dados brutos de entrada em um formato apropriado para análise. As principais funções do pré-processamento são descritas da seguinte maneira:

1. **Seleção dos dados:** essa função, também chamada de Redução de Dados, envolve a identificação de quais informações, dentre as bases de dados existentes, devem ser realmente consideradas durante o processo de KDD. Deve ser considerado no processo o aspecto dos dados, que pode ser de atributos ou de registros. (Goldschmidt e Passos, 2005);

2. **Limpeza de dados:** de maneira frequente, os dados são encontrados com diversas inconsistências como, por exemplo, registros incompletos, valores errados e dados inconsistentes. Com isso, essa tarefa consiste na remoção desse ruídos. (Tan et al., 2014);

3. **Codificação dos dados:** essa etapa consiste em codificar os dados de tal maneira que possam ser usados como entrada para os algoritmos de mineração de dados, por exemplo, quando se transforma uma variável numérica em categórica, em

que os valores reais são transformados em intervalos (Goldschmidt e Passos, 2005).

4. **Enriquecimento dos dados:** o enriquecimento se constitui em conseguir, de alguma maneira, mais informação que possam ser agregadas aos registros existentes, enriquecendo os dados, para que eles possam fornecer mais informações no processo KDD. Para isso, podem ser realizadas novas pesquisas complementares, consulta à base de dados externas, entre outras técnicas (Goldschmidt e Passos, 2005).

Mineração de dados

De acordo com Witten et al. (2011b), mineração de dados consiste no processo da descoberta automática de informações implícitas, potencialmente úteis dos dados, de tal modo que se revelem regularidades ou padrões nos dados, podendo ser generalizados para fazer, por exemplo, predições sobre dados futuros.

Para Goldschmidt e Passos (2005), dentro do KDD, o processo de mineração de dados requer a definição das técnicas e dos algoritmos que serão utilizados no problema em questão. A escolha da técnica depende, muitas vezes, do tipo de tarefa a ser realizada.

Estas tarefas geralmente são divididas em duas categorias principais: tarefas para mineração de dados preditiva e mineração de dados descritiva. A tarefa preditiva tem como objetivo prever o valor de um determinado atributo com bases em valores de outras variáveis. Já nas tarefas descritivas, o objetivo é derivar padrões (correlações, tendências, *clusters*, trajetórias e anomalias) que resumem o relacionamento dos dados (Tan et al., 2014).

A mineração de dados inclui inúmeras tarefas, como: classificação, regressão, associação e agrupamento. Essas tarefas podem descobrir diferentes tipos de conhecimentos e são brevemente descritas abaixo:

1. **Classificação:** a classificação consiste na tarefa de aprender uma função que mapeie um conjunto de registros em um conjunto de rótulos categóricos predefinidos, denominados classes. Uma vez que se descobre tal função, ela pode ser aplicada a novos registros de forma a prever a classe em que tais registros se enquadram (Goldschmidt e Passos, 2005). A classificação pode ser descritiva quando ela serve como ferramenta explicativa para diferenciar classes e objetos diferentes. Ou pode ser preditiva, que prevê o rótulo de classe de registros não conhecidos com base nos dados coletados no passado.

2. **Regressão:** a regressão consiste no aprendizado de uma função que mapeie registros de um banco de dados em valores reais. Esta tarefa é similar à tarefa de classificação, sendo restrita apenas a atributos numéricos (Fayyad et al. 1996; Goldschmidt e Passos 2005). Compreende a busca por uma função que mapeie os registros de um banco de dados em valores reais.

3. **Regras de Associação:** as regras de associação enfatizam a análise das relações entre as variáveis, e não entre os objetos de uma base de dados. Esse tipo de análise costuma ser utilizado em ações de *marketing* e para o estudo de bases de dados transacionais. Dessa maneira, um bom algoritmo de mineração de regras de associação precisa ser capaz de estabelecer associações entre itens que sejam estatisticamente relevantes para o universo representado pela base de dados (Ferrari e De Castro, 2017).

4. **Agrupamento:** utilizado para separar os registros de uma base de dados em subconjuntos (*clusters*), de tal maneira que os elementos de um *cluster* compartilham propriedades comuns que os distingam de elementos de outros *clusters* (Goldschmidt e Passos, 2005). Métodos de agrupamento ou *clustering* são utilizados para dividir objetos de dados em grupos, ou então, como um passo de pré-processamento para outros algoritmos (Tan et al., 2014). Diferente da tarefa de classificação, que se baseia em rótulos predefinidos, a clusterização precisa automaticamente identificar os grupos de dados (Fayyad et al., 1996).

2.5 Mineração de Dados e Aprendizado de Máquina

Diversos conceitos da área de Aprendizado de Máquina acabaram favorecendo o campo de Mineração de Dados. Muitos dos conceitos vistos neste capítulo são abordados em livros voltados para a área de Aprendizado de Máquina, como, por exemplo, as técnicas de Classificação, Regressão, Regras de Associação e Agrupamento. Os dois grupos de Mineração de Dados, preditiva e descritiva, são análogos à forma em que os tipos de Aprendizado são fragmentados: Aprendizado Supervisionado e Não-Supervisionado. Witten et al. (2011a) estabelece que o aprendizado de máquina fornece base técnica para a mineração de dados.

2.5.1 Aprendizado de Máquina

Aprendizado de Máquina consiste em uma abordagem guiada para a resolução de inúmeros problemas, na qual padrões em um conjunto de dados são identificados e utilizados para apoiar nas tomadas de decisões (Brink et al., 2016).

O Aprendizado de Máquina está diretamente ligado à Mineração de Dados e à Estatística, mas foca nas propriedades dos métodos estatísticos, assim como sua complexidade computacional. Na prática, Aprendizado de Máquina é uma subárea da Inteligência Artificial que busca, por meio da construção de modelos matemáticos, mediante um conjunto de dados que descrevem um fenômeno (conjunto de treinamento), realizar predições acerca de um assunto de interesse (Bishop, 2006). Uma das grandes vantagens do aprendizado de máquina sobre os métodos clássicos de predição é sua capacidade em lidar com bases de dados complexas que incluem dados quantitativos e qualitativos.

As principais abordagens de aprendizado normalmente aplicadas em Mineração de Dados são: aprendizado supervisionado e aprendizado não supervisionado

No Aprendizado Supervisionado, para cada exemplo apresentado ao algoritmo de aprendizado é necessário apresentar um rótulo especificando a que classe o exemplo pertence. Cada exemplo é composto por um conjunto de atributos e pelo rótulo da classe associada. O objetivo do algoritmo é construir um classificador que possa determinar de maneira correta a classe de novos exemplos ainda não rotulados. Para rótulos de classe discretos, esse problema é chamado de classificação e para valores contínuos como regressão (Ludermir, 2021).

No Aprendizado Não Supervisionado, os exemplos de entrada não possuem rótulos de classe. O algoritmo agrupa os exemplos pelas semelhanças dos seus atributos. O algoritmo analisa os exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados de alguma maneira, formando agrupamentos ou *clusters*. Após a determinação dos agrupamentos, em geral, é necessária uma análise para determinar o que cada agrupamento significa no contexto do problema analisado (Ludermir, 2021).

2.6 Algoritmos de Aprendizado de Máquina

2.6.1 Árvores de Decisão

As Árvores de Decisão (AD) constituem métodos que utilizam-se de uma representação com base em árvores. Para a construção da árvore é feita uma sequência de divisões no conjunto de dados baseadas na seleção de melhores pontos de separação de valores dos atributos e, assim, determina-se a ramificação da árvore até que um determinado critério de parada seja satisfeito. Existe o nó raiz, em que inicia a árvore; os nós de decisão, que estabelecem o caminho da árvore; e os nós terminais/folha, que representam uma classe ou valor contínuo Breiman et al. (1984).

Os modelos em árvore são designados de decisão no caso de problemas de classificação e árvore de regressão nos problemas de regressão. A interpretação dos modelos e dos algoritmos de árvores de decisão e regressão são muito semelhantes, isso porque os dois tipos de árvores são formados por um conjunto de nós de decisão. A diferença é que o resultado da árvore de decisão (classificação) é uma categoria, enquanto na de regressão é um escalar.

De acordo com Loh (2011), as árvores de classificação são implementadas para variáveis dependentes que recebem um número finito de valores, com erro de previsão medido em nível de custo de classificação incorreta. Já as árvores de regressão são projetadas para variáveis dependentes que aceitam valores discretos, contínuos ou ordenados, com erro de previsão medido por meio da diferença quadrática entre os valores observados e preditos.

Segundo Faceli et al. (2011), pode-se definir formalmente a árvore de decisão como

um grafo acíclico direcionado em que cada nó ou é um *nó de divisão*, com dois ou mais sucessores, ou um *nó folha*:

- Um *nó folha* é considerado um nó objetivo ou nó externo. No percurso de uma árvore por meio dos nós de decisão, os nós folhas representam os resultados da predição.
- Um *nó decisão* possui um teste condicional baseado nos valores dos atributo, usado para se caminhar na árvore até atingir as folhas. O primeiro nó de uma árvore é denominado de nó raiz, e este é considerado um nó decisão.

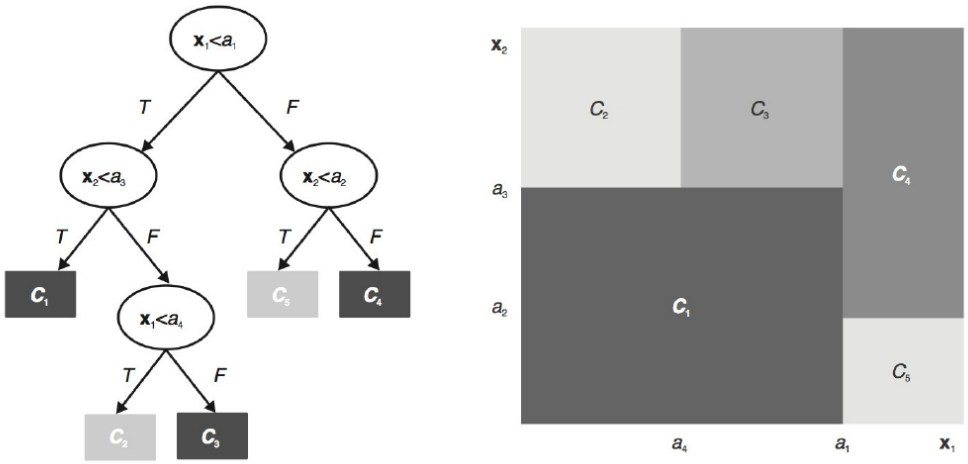

Figura 2.5: Uma árvore de decisão e as regiões de decisão no espaço de objetos (Faceli et al., 2011).

Com o objetivo de proporcionar uma melhor entendimento sobre as árvores de decisão, a Figura 2.5 ilustra um exemplo citado por Faceli et al. (2011) na qual foi criada uma árvore de decisão para representar a divisão correspondente no espaço pelos atributos X_1 e X_2. Cada nó da árvore corresponde a uma região nesse espaço. As regiões definidas pelas folhas da árvore são mutuamente excludentes, e a reunião dessas regiões cobre todo o espaço definido pelos atributos. A interseção das regiões abrangidas por quaisquer duas folhas é vazia. Observa-se que a árvore da Figura 2.5 possui cinco folhas, logo haverá cinco regras de decisão para essa árvore. Ao aplicar a regra SE-ENTÃO lê-se a árvore da seguinte forma:

1. Se X_1 menor que a_1 e X_2 menor que a_3, então o espaço correspondente é C_1;
2. Se X_1 não é menor que a_1 e X_2 é menor que a_2, então o espaço correspondente é C_5;
3. Se X_1 não é menor que a_1 e X_2 não é menor que a_2, então o espaço correspondente é C_4;
4. Se X_1 é menor que a_1 e X_2 não é menor que a_3 e X_1 é menor que a_4, então o espaço correspondente é C_2;

5. Se X_1 é menor que a_1 e X_2 não é menor que a_3 e X_1 não é menor que a_4, então o espaço correspondente é C_2;

Uma árvore de decisão engloba todo o espaço de instâncias. Esse motivo implica que uma árvore de decisão pode fazer predições para qualquer exemplo de entrada.

2.6.2 *Random Forest*

A Random Forest (RF) é um algoritmo do tipo *ensemble* utilizado para classificação e regressão que combina o resultado de um conjunto de árvores de decisão para realizar a predição. Cada uma dessas árvores de decisão é construída utilizando uma amostra aleatória inicial dos dados e, a cada divisão desses dados, um subconjunto aleatório de m atributos é utilizado para escolha do atributo mais importante. Para isso, utiliza a estratégia denominada como *Bagging* (um meta-algoritmo para melhorar a classificação e a regressão de modelos de acordo com a estabilidade e a precisão da classificação) para construir suas árvores, e dessa forma, cria uma floresta com baixa correlação (Breiman, 2001).

A RF utiliza a ideia de que a decisão por voto para predição final é melhor do que a previsão individual de cada árvore (Breiman, 2001). O modelo final de RF elabora uma listagem das variáveis mais importantes na construção da floresta, que são determinados pela importância acumulada da variável nas divisões dos nós de cada árvore da floresta (James et al., 2013).

De maneira simplificada, Breiman (2004) apresenta o funcionamento de uma RF acompanhando o seguinte sumário:

1. Os dados utilizados no treinamento são uma amostra da base de dados.
2. Um número N de variáveis disponíveis para divisão em cada nó da árvore é definido pelo usuário, onde valor N é menor que o número total de variáveis.
3. Em cada nó as N variáveis são selecionadas por amostragem aleatória, independente e uniforme, e o nó é dividido na melhor divisão possível considerando as variáveis selecionadas.
4. Depois da seleção das variáveis, a RF determina a importância dos atributos de cada árvore através de um índice, tais como tais como impureza, distância e dependência. (Breiman, 2001). A partir desses índices é possível estimar a distribuição das classes do atributo em cada nó. A divisão de cada nó é feita de forma a gerar nós filhos mais "puros" do que o nó anterior, ou seja, com maiores concentrações de exemplos de uma mesma classe.

A Figura 2.6 ilustra a estrutura de classificação de uma RF, onde a previsão final é baseada na classe que obtiver o maior número de votos. No caso da regressão, como um vetor de teste é adicionado a cada árvore, é atribuído o valor médio dos valores do nó terminal.

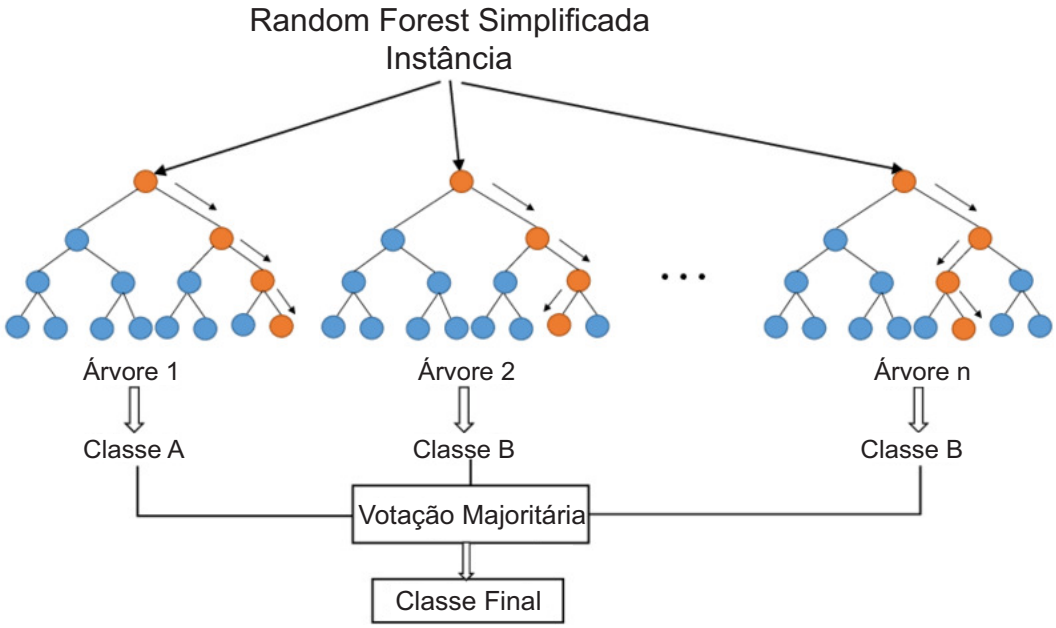

Figura 2.6: *Estrutura de Classificação da Random Forest* (Tran, 2019)

O RF é uma algoritmo eficaz na previsão, com menos tendência a *overfit* (Breiman, 2001), devido à aleatoriedade na seleção de atributos entre as árvores. Além disso, a RF é considerada como um algoritmo muito útil e fácil de utilizar, visto que os hiperparâmetros padrão, muitas vezes, permitem produzir um bom resultado. Contudo, uma previsão com maior precisão requer mais árvores, o que pode resultar em um modelo mais lento (Tran, 2019).

2.6.3 *Light Gradient Boosting Machines*

O *Light Gradient Boosting Machines* (LGBM) é um *Gradient Boosting Framework* (GBF) baseado em árvores de decisão. A essência da concepção de um GBF consiste em um procedimento que combina a saída preditiva de inúmeros classificadores que são considerados "fracos". Cada modelo "fraco" adicional vai diminuindo o erro quadrático médio do modelo geral com o objetivo de conceber um comitê poderoso, responsável pela predição final. Geralmente, esses modelos são árvores de decisão (Hastie et al., 2009; James et al., 2013).

Diferente dos outros modelos baseados em árvores, no LGBM, as árvores crescem na vertical, ou seja, utilizam a estratégia de crescimento de árvore em folha, enquanto os outros algoritmos de *boosting* crescem a árvore horizontalmente, isto é, crescem a profundidade da árvore em nível. O LGBM seleciona a folha da árvore com o melhor ajuste para crescer e assim pode diminuir mais perdas do que um algoritmo com base no crescimento horizontal (Ke et al., 2017). As Figuras 2.7 e 2.8 ilustram as estratégias de crescimento de árvores em nível e em folha, respectivamente.

Figura 2.7: Estratégia de crescimento de árvores em nível (Microsoft, 2022)

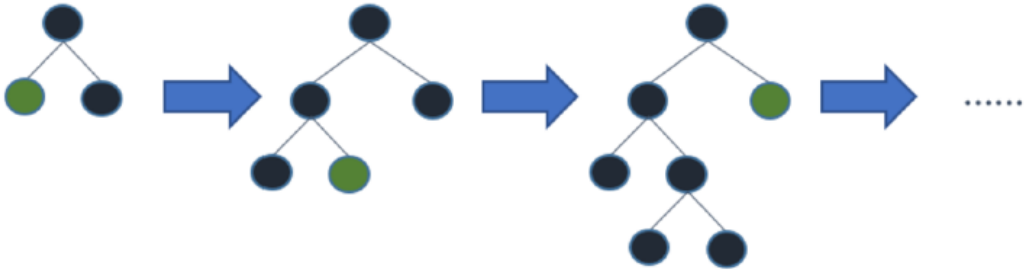

Figura 2.8: Estratégia de crescimento de árvores em folha (Microsoft, 2022)

O algoritmo LGBM utiliza duas técnicas principais: *Gradient-based One-Side Sampling* (GOSS) e *Exclusive Feature Bundling* (EFB). O GOSS possibilita a exclusão de uma proporção significativa de instâncias de dados com gradientes pequenos, utilizando as demais para estimar o ganho de informação, e, com isso, mantendo uma estimativa precisa de ganho de informação com um conjunto de dados muito menor. O EFB possibilita agrupar as variáveis mutuamente exclusivas, com o objetivo de diminuir a dimensionalidade dos dados. Encontrar um agrupamento ideal das variáveis exclusivas é um problema NP-difícil, contudo a utilização de um algoritmo guloso permite ao LGBM alcançar boa taxa de aproximação. Desta forma, o LGBM consegue acelerar o processo de treinamento do *gradient boosting* baseado em árvore de decisão, em até 20 vezes do que outros algoritmos que utilizam GBF, sem perder precisão (Ke et al., 2017).

2.6.4 Redes Neurais Artificiais

Redes Neurais Artificiais (RNA) surgiram na década de 1940, assim como o termo Inteligência Artificial. Estas foram desenvolvidas com o objetivo de simular a estrutura e o funcionamento do cérebro humano, de maneira que consiga adquirir algumas de suas características, tais como: capacidade adaptativa, velocidade de processamento, processamento paralelo e aprendizado (Haykin, 2009).

De acordo com Haykin (2009), uma RNA pode ser entendida, como um arranjo de unidades, neurônios interconectados, trabalhando de forma paralela para classificar

e generalizar dados de entrada em classes de saída. Um neurônio consiste em uma unidade de processamento de informação fundamental para o funcionamento de uma rede neural. Na Figura 2.9 é apresentado um modelo de neurônio artificial, que é a base para implementar diversos tipos de RNAs. Pode-se destacar três elementos básicos de um modelo neural: (i) um conjunto de *pesos sinápticos* que podem incluir valores negativos e positivos; (ii) um *somador*, que soma as entradas ponderadas pelos respectivos pesos sinápicos; (iii) uma *função de ativação* que tem o objetivo de limitar a amplitude da saída para algum valor finito. O modelo da Figura 2.9 também inclui um *bias* que tem o efeito de aumentar ou diminuir a entrada da função de ativação.

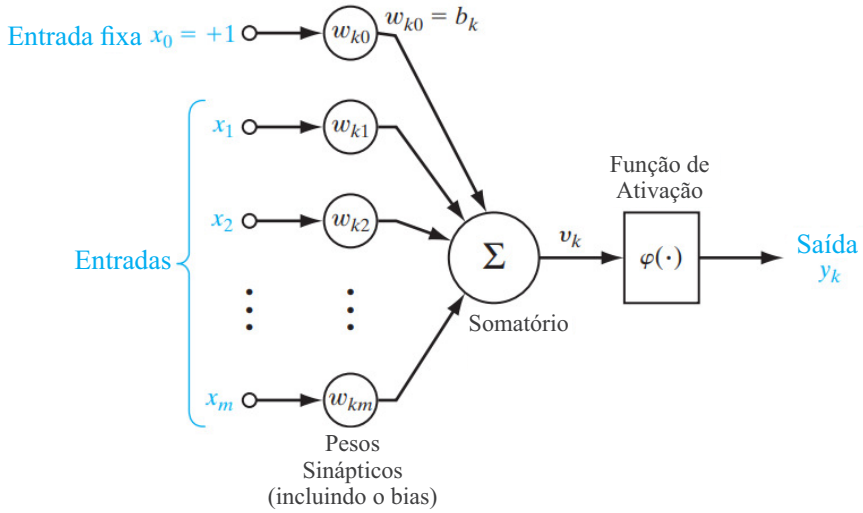

Figura 2.9: Modelo não linear de um neurônio, rotulado k (Haykin, 2009).

Matematicamente, pode-se descrever o neurônio k representado na Figura 2.9 pelo par de equações:

$$v_k = \sum_{j=1}^{m} w_{kj} x_j \tag{2.1}$$

e

$$y_k = \varphi(v_k) \tag{2.2}$$

onde $x_1, x_2, ..., x_m$ são os sinais de entrada; $w_{k1}, w_{k2}, ..., w_{km}$ são os respectivos pesos sinápticos do neurônio k; x_o é uma entrada com valor de $+1$; w_{k0} é igual ao bias b_k que é um parâmetro externo do neurônio, podendo ser positivo ou negativo; v_k é a saída do combinador linear devido aos sinais de entrada; $\varphi(\cdot)$ é a função de ativação; e y_k é o sinal de saída do neurônio.

Conforme Haykin (2009) e Faceli et al. (2011), uma função de ativação, denotada por $\varphi(v)$, define a saída de um neurônio em termos de campo local v. Várias funções

de ativação têm sido propostas na literatura. A Figura 2.10 mostra a equação e o formato de duas dessas funções, as funções limiar e sigmoide. Na função limiar (Figura 2.10 (a)), o valor do limiar define quando o resultado da função limiar será igual a 1 ou 0. Quando o somatório das entradas recebidas supera o limiar definido, o neurônio torna-se ativo (saída +1). Quanto maior é o valor do limiar, maior tem que ser o valor da entrada total para que o valor de saída do neurônio seja igual a 1. Na função sigmoide (Figura 2.10 (b)), o gráfico tem forma de "S", sendo a função de ativação mais empregada na construção de redes neurais. É definida como uma função estritamente crescente que exibe o equilíbrio entre o comportamento linear e não linear e representa uma aproximação contínua e diferenciável da função limiar. Um exemplo da função sigmoide é a função logística, definida por:

$$\varphi(v) = \frac{1}{1 + \exp(-av)} \tag{2.3}$$

onde a é o parâmetro de inclinação da função sigmoide.

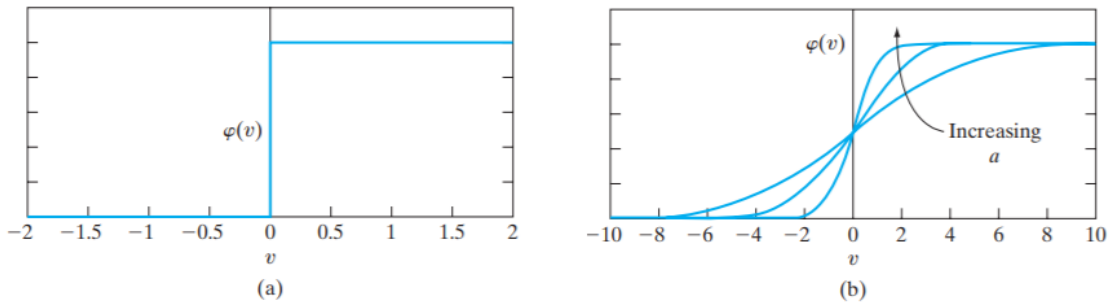

Figura 2.10: (a) Função de limiar. (b) Função sigmoide (Haykin, 2009).

Ao longo do tempo, foram desenvolvidos um grande número de arquiteturas de RNAs, bem como algoritmos de treinamentos. Serão apresentadas algumas das principais arquiteturas existentes e os algoritmos de treinamento mais utilizados, com foco em modelos de paradigma de aprendizado supervisionado.

Redes Perceptron e Adaline

De acordo com Haykin (2009), a rede *Perceptron* foi a primeira RNA implementada e foi construída em torno de um neurônio não linear. Embora seja uma rede simples, apresentando apenas uma camada de neurônios, ela apresentou boa acurácia preditiva em alguns problemas de classificação.

Conforme Faceli et al. (2011), a rede é treinada por meio de um algoritmo supervisionado de correção de erro e utiliza a função de ativação do tipo limiar. Em seu treinamento, os pesos são ajustados, para um objeto x_i, conforme a equação 2.4.

$$w_j(t+1) = w_j(t) + \eta x_i^j (y_i - \hat{f}(x_i)) \tag{2.4}$$

onde $w_j(t)$ é o peso da j-ésima conexão de entrada no instante de tempo t, η é uma taxa de aprendizado, x_i^j é o valor do j-ésimo atributo do vetor de entrada x_i, $\hat{f}(x_i)$ é a saída produzida pela RNA no instante t e y_i é a saída desejada para a rede.

A rede *Adaline* surgiu nessa mesma época e a principal diferença em relação à *Perceptron* é que ela utiliza uma função de ativação linear e, desta forma, leva em consideração a magnitude do erro em consideração no momento de ajustar os pesos da rede. Para tanto, a rede *Adaline* utiliza uma regra denominada Regra Delta, cujo nome vem da diferença entre os valores da saída desejada e da saída produzida. O valor da saída da *Adaline* é definido como contínuo, com isso, são redes comumente utilizadas para problemas de regressão. Para problemas de classificação, as saídas dos neurônios devem ser discretizadas. As redes *Perceptron*, por outro lado, foram desenvolvidas para a solução de problemas de classificação. Uma limitação das redes *Perceptron* e *Adaline* é que só conseguem classificar objetos que são linearmente separáveis (Faceli et al., 2011).

Multilayer Perceptrons

Tanto o *Perceptron* quanto o *Adaline* não foram capazes de resolver problemas não linearmente separáveis, para isso, a alternativa mais utilizada é adicionar uma ou mais camadas. Com isso, olhamos para uma estrutura conhecida como *multilayer Perceptron* ou *Perceptron* multicamadas - MLP (Figura 2.11). Haykin (2009) destaca três características básicas dos MLPs, a saber:

1. Utilizam uma função de ativação não linear que é diferenciável;
2. A rede é constituída de uma ou mais camadas ocultas entre a camada de entrada e saída;
3. Possui alto grau de conectividade, cuja extensão é determinada pelos pesos sinápticos.

A Figura 2.11 mostra um exemplo de uma rede MLP com duas camadas ocultas e exibe o papel desempenhado pelos neurônios em cada uma das camadas. Na primeira camada, cada neurônio aprende uma tarefa que define um subespaço, em que separa o espaço das variáveis de entrada em duas partes. Cada um dos neurônios da camada seguinte realiza a combinação de um grupo de subespaços que foi definido pelos neurônios da camada anterior, gerando regiões convexas. Os neurônios da camada seguinte realizam a combinação de um subconjunto das regiões convexas em regiões de formato qualquer. Desta forma, a combinação das funções desempenhadas por cada um dos neurônios da rede, vai determinar a função associada à rede como um todo (Faceli et al., 2011).

O erro cometido pela rede para a classificação de um determinado objeto é definido pela comparação entre o vetor de saída dos neurônios na camada de saída e o vetor de

valores desejados para essas saídas. Um objeto é classificado corretamente quando o valor de saída mais elevado produzido pela rede é aquele gerado pelo neurônio de saída que corresponde à classe correta do objeto, já o erro ocorre quando o neurônio de outra classe produz o valor mais elevado. Quando nenhum valor elevado é gerado ou quando muitos neurônios produzem valores elevados a rede não consegue prever a classe do objeto. No caso de problemas de regressão não se tem a discretização imposta pela escolha do neurônios com maior saída na predição (Faceli et al., 2011).

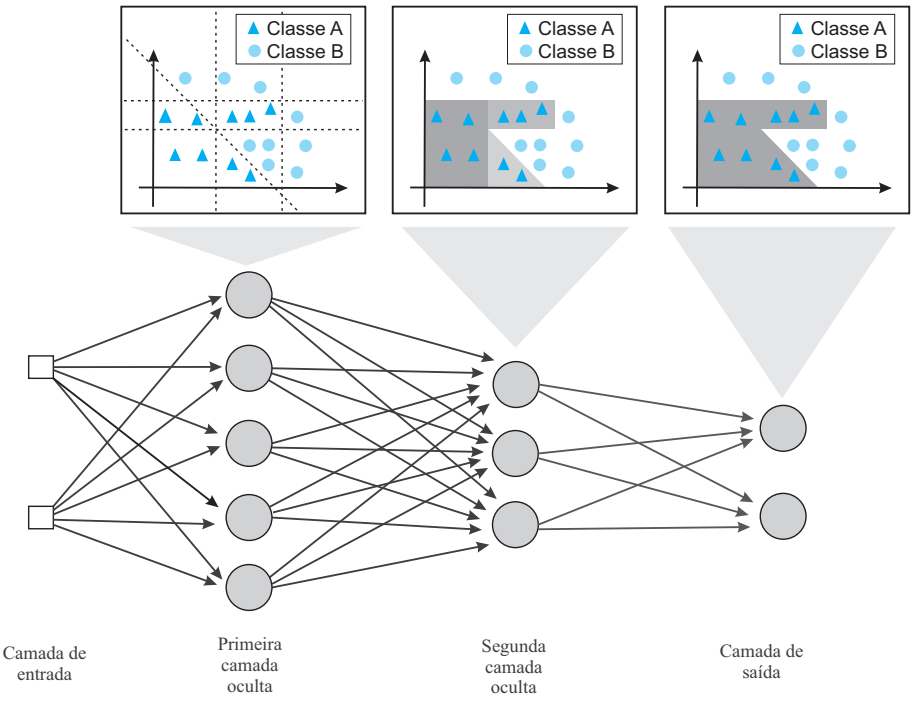

Figura 2.11: Papel desempenhado pelos neurônios das diferentes camadas da rede MLP. Adaptado. (Faceli et al., 2011)

Algoritmo Back-propagation

O algoritmo *back-propagation*, baseado em gradiente descendente, possibilitou que as MLPs apresentassem capacidade de aprendizado, sendo necessária a utilização de uma função de ativação contínua, diferenciável e, de preferência, não decrescente. Ele retro-propaga o valor do erro de cada neurônio da camada de saída que é utilizado para ajustar seus pesos de entrada. O ajuste ocorre da camada de saída até a primeira camada oculta da rede (Faceli et al., 2011). A Equação 2.5 mostra como é realizado o ajuste dos pesos de uma rede MLP pelo algoritmo *back-propagation*.

$$w_{jl}(t+1) = w_{jl}(t) + \eta x^j \delta_l \tag{2.5}$$

onde w_{jl} corresponde ao peso entre um neurônio l e o j-ésimo atributo de entrada ou a saída do j-ésimo neurônio da camada anterior, δ_l indica o erro associado ao

l-ésimo neurônio e x^j indica a entrada recebida por esse neurônio (o j-ésimo atributo de entrada ou a saída do j-ésimo neurônio da camada anterior).

O erro de um neurônio de uma determinada camada oculta é estimado como a soma dos erros dos neurônios da camada seguinte, no qual as extremidade de entrada estão ligadas a ele, ponderados pelo valor do peso associado a essas conexões. Dessa forma, o cálculo do erro vai depender da camada em que o neurônio se encontra, conforme a equação apresentada por (Faceli et al., 2011):

$$\delta_l = \begin{cases} f'_a e_l, & \text{Se } n_l \in c_{sai} \\ f'_a \sum w_{lk}\delta_k, & \text{Se } n_l \in c_{ocu} \end{cases} \tag{2.6}$$

onde n_l é o l-ésimo neurônio, c_{sai} consiste na camada de saída, c_{ocu} representa uma camada oculta, f'_a é a derivada parcial da função de ativação do neurônio e e_l é o erro quadrático produzido pelo neurônio de saída quando sua resposta é comparada à desejada. O erro quadrático é definido pela Equação 2.7.

$$e_l = \frac{1}{2}\sum_{q=1}^{k}(y_q - \hat{f}_q)^2 \tag{2.7}$$

onde y_q é a saída desejada e $\hat{f}_q$ a saída produzida pelo neurônio.

2.6.5 Otimização dos classificadores através de hiperparâmetros

Os classificadores possuem um conjunto de parâmetros que irão conduzi-los no processo de aprendizado e classificação, que influenciam na qualidade do modelo de acordo com a métrica escolhida. Com o objetivo de melhorar o desempenho do modelo, faz-se necessário ajustar as configurações de parâmetros a fim de encontrar valores ótimos para cada um deles. Esse processo pode ser exaustivo, devido à grande quantidade de parâmetros a serem alterados e as possibilidades de configuração de cada classificador. Dentre as estratégias mais utilizadas para otimização dos hiperparâmetros estão o *Grid Search* e o *Random Search*. As duas estratégias são descritas a seguir.

Grid Search: Trata-se de uma pesquisa em grade, que de maneira simples faz uma pesquisa completa sobre determinado subconjunto do espaço de hiperparâmetros do algoritmo de treinamento. Como o hiperparâmetro do classificador pode incluir espaço com valores ilimitados, é possível que seja necessário especificar um limite para aplicar a *grid search* (Bergstra e Bengio, 2012).

Random Search: Ele substitui a estratégia de busca na seleção completa de combinações por um busca aleatória em um determinado subconjunto. Isso pode ser facilmente aplicado a casos discretos, mas o método pode ser generalizado para espaços contínuos e mistos. A pesquisa aleatória pode superar o *grid search*, especialmente se apenas um pequeno número de parâmetros afetar o desempenho do

classificador. No entanto, apesar de levar menor tempo para encontrar o melhor resultado, tende a ser menos preciso que o *grid search* (Bergstra e Bengio, 2012).

2.7 Treinamento e Teste

De forma a construir e avaliar o desempenho dos algoritmos de classificação, os mesmos são submetidos a pelo menos dois conjuntos de dados, que são recortes da base de dados originais: um conjunto para treinamento e outro conjunto para teste. A utilização do conjunto de teste se dará no final do processo, apenas após as etapas de otimização dos hiperparâmetros e validação dos modelos.

Durante a etapa de treinamento, os modelos utilizam os atributos para determinar a classe alvo e comparar o resultado obtido com a classe alvo já conhecida para realizar o ajuste do modelo do classificador. Esse tipo de procedimento de treinamento é chamado de treinamento supervisionado, visto que utiliza a informação conhecida da classe alvo para acertar a classificação.

Ao término do treinamento, o modelo de classificação é apresentado ao conjunto de testes com dados que ainda não foram utilizados na avaliação do modelo, de modo a estimar a capacidade do modelo em dados inéditos. Com isso, os resultados são analisados, permitindo a avaliação do desempenho do modelo na predição em dados gerais (Zaki et al., 2014).

2.8 Protocolos de Validação

De acordo com Kohavi (1995), com o objetivo de avaliar a desempenho de predição, utilizando para teste parte dos dados separados dos dados usados na fase de treinamento, são utilizados técnicas como *Hold-out* e *Cross-Validation*.

2.8.1 *Hold-out*

Han et al. (2011), elucida que o método *Hold-out* consiste em dividir os dados em dois conjuntos independentes: um conjunto de treinamento e um conjunto de teste. Essa divisão é feita com o objetivo de avaliar o desempenho da predição em dados diferentes dos usados no treinamento.

O *Hold-out* realiza uma estimativa limitada do desempenho do classificador, pois utiliza apenas um conjunto de treinamento e teste. Com o intuito de minimizar as limitações do *Hold-out*, pode-se aplicar a subamostragem aleatória, que consiste em uma variação do método *Hold-out* no qual esse método é repetido n vezes. A estimativa de desempenho geral é tomada como a média da desempenho de cada iteração (Han et al., 2011).

2.8.2 *K-Fold Cross-Validation*

De acordo com Kohavi (1995), o método de *k-Fold Cross-Validation* consiste em dividir aleatoriamente um conjunto de dados D em k subconjuntos mutuamente exclusivos de tamanho aproximadamente igual.

Han et al. (2011) estabelece os seguintes passos para execução do *k-Fold Cross-Validation*:

1. Os dados iniciais são divididos aleatoriamente em k subconjuntos mutuamente exclusivos ou "*folds*", f1, f2,..., fk , cada um de tamanho aproximadamente igual;
2. Na primeira iteração, os subconjuntos f2,..., fk consistem no conjunto de treinamento para obter um primeiro modelo, que é testado em f1;
3. A segunda iteração é treinada nos subconjuntos f1, f3,..., fk e testado em f2;
4. O processo é repetido até que o número de iterações ser igual a k.

A Figura 2.12, ilustra o processo do *k-fold Cross-Validation*. Os dados de treinamento foram divididos em k partes, e em cada iteração, k-1 partes foram utilizadas para o treinamento do modelo com diferentes hiperparâmetros, e uma parte para estimar sua performance preditiva. Ao término do processo, as performances estimadas em cada modelo, são utilizadas para calcular a performance média.

A repetição em k vezes procura reduzir o viés, já que está sendo usado a maioria dos dados para ajuste, o que também reduz significativamente a variação, pois a maioria dos dados também está sendo usada no conjunto de testes.

Repetições do *Cross-Validation* podem ser realizadas com o objetivo de aumentar a precisão das estimativas de desempenho (Kuhn e Johnson, 2013). No *stratified cross-validation*, os *folds* são estratificados para que a distribuição de classes em cada *fold* seja aproximadamente a mesma que nos dados originais (Han et al., 2011).

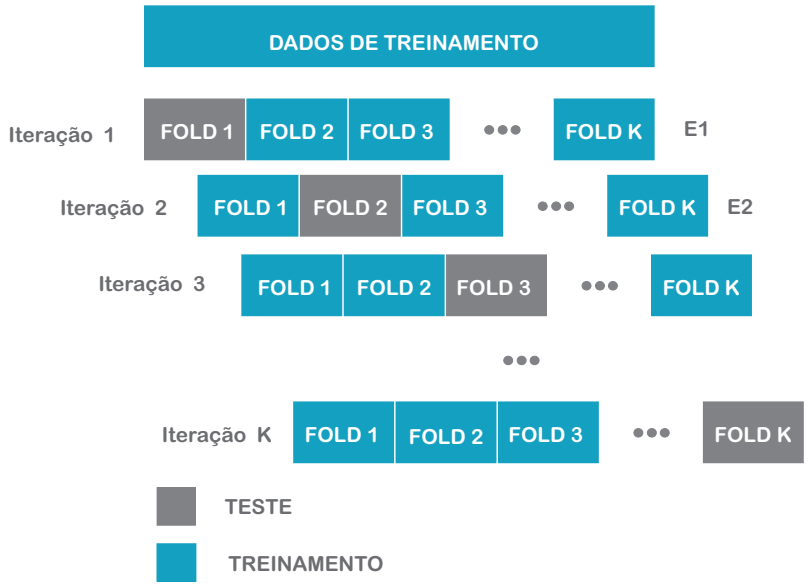

Figura 2.12: Processo do *Cross-Validation*.

2.9 Medidas de Avaliação

Após a construção dos classificadores, é necessário que eles sejam avaliados quantitativamente. As medidas utilizadas têm grande impacto de como um determinado classificador pode ser compreendido. Dessa maneira, a escolha da medida avaliativa faz parte do processo de mineração de dados (Kohavi e Provost, 1998). Nas próximas seções serão apresentadas as medidas de avaliação utilizadas nesta pesquisa.

2.9.1 Matriz de Confusão

A matriz de confusão pode ser utilizada como premissa para uma série de técnicas estatísticas e analíticas. Ela mostra o número de elementos que foram corretamente ou incorretamente classificados para cada classe. Na Tabela 2.6, pode-se verificar, na diagonal principal, o número de instâncias que foram corretamente classificadas para cada classe, enquanto que os elementos fora da diagonal principal indicam o número de amostras que foram classificados incorretamente (Maimon e Rokach, 2010).

Tabela 2.6: Matriz de Confusão para problemas com duas classes.

	Predição Positiva	Predição Negativa
Classe Positiva	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Classe Negativa	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Para realizar o cálculo de algumas métricas de avaliação, faz-se necessário compreender alguns termos utilizados na matriz de confusão (Han et al., 2012).

- Verdadeiros positivos (VP): são as amostras positivas que foram corretamente rotuladas pelo classificador;
- Verdadeiros negativos (VN): são as amostras negativas que foram corretamente rotuladas pelo classificador;
- Falsos positivos (FP): são as amostras negativas que foram incorretamente rotuladas como amostras positivas pelo classificador;
- Falsos negativos (FN): são as amostras positivas que foram incorretamente rotuladas como amostras negativas pelo classificador.

2.9.2 Acurácia

Acurácia é o percentual de casos que foram classificados corretamente, sem levar em consideração o que é positivo e o que é negativo, ou seja, é a taxa de acerto global. Sua fórmula é representada pela equação 2.8.

$$ACC = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.8}$$

2.9.3 Precisão

A precisão diz respeito à proporção de amostras classificadas corretamente como positivas. A Equação 2.9 mostra como obter o valor da precisão.

$$P = \frac{VP}{VP + FP} \tag{2.9}$$

2.9.4 Sensibilidade ou *Recall*

A Sensibilidade (S) de um classificador indica a proporção de casos positivos que foram identificados corretamente. O cálculo da sensibilidade é feito com base na matriz de confusão, seu cálculo é realizado conforme a Equação 2.10.

$$S = \frac{VP}{VP + FN} \tag{2.10}$$

2.9.5 Especificidade

A Especificidade (E) de um classificador reflete a proporção de casos negativos que foram identificados corretamente, sua fórmula corresponde à equação 2.11.

$$E = \frac{VN}{VN + FP} \tag{2.11}$$

2.9.6 Medida F1

A medida F1 leva em consideração tanto a precisão quanto a sensibilidade. Essa medida é estabelecida pela média harmônica entre as duas, como pode ser visto na Equação 2.12.

$$F1 = 2 * \frac{Precisão * Sensibilidade}{Precisão + Sensibilidade} \tag{2.12}$$

2.9.7 Curva característica de Operação do Receptor - ROC e Área sobre a curva - AUC

Uma ferramenta que pode ser aplicada em problemas binários de classificação é a área sobre a curva ROC (*Receiver Operating Characteristic*), visto que por meio dela pode se realizar medidas de desempenho independentes de condições como limiar de classificação, assim como custos associados às classificações incorretas e à distribuição de classes (Faceli et al., 2011).

A curva ROC possui grande utilidade na avaliação e comparação entre algoritmos ou quando é necessário levar em consideração diferentes custos/benefícios para os diferentes erros/acertos em um processo de classificação, visto que denota os acertos para as duas classes em análise (Prati et al., 2008).

O gráfico ROC é bidimensional plotado em um espaço designado espaço ROC, com eixos X e Y representando as medidas de taxa de falsos positivos (TFP) e taxa de verdadeiros positivos (TVP), respectivamente. A Figura 2.13 apresenta o espaço ROC, apenas com classificadores discretos, no qual a linha diagonal representa os classificadores que realizam predições aleatórias. Classificadores discretos são aqueles que geram como saída somente uma classe. Estes classificadores fornecem um par (Taxa de FP, Taxa de VP) correspondendo a um ponto no espaço ROC. Desta forma, o ponto (0,1), denominado de céu ROC, representaria o melhor método de previsão possível, visto que nesse ponto todas as amostras positivas e negativas são classificados corretamente, ou seja, não teríamos falsos negativos e não teríamos falsos positivos (Faceli et al., 2011).

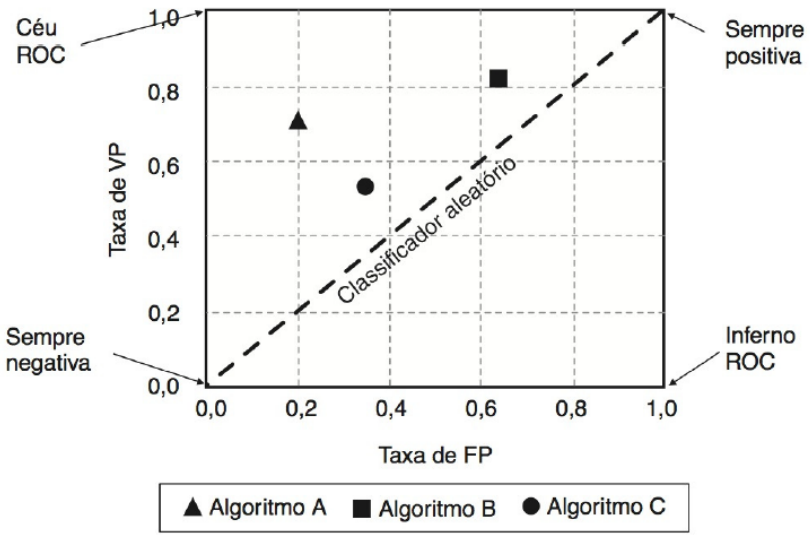

Figura 2.13: Espaço ROC (Faceli et al., 2011).

Existem alguns mecanismos de se comparar diferentes classificadores com base nos pontos do espaço ROC, sendo o mais usual gerar uma curva ROC. A Figura 2.14 ilustra duas curvas ROC, que de forma hipotética correspondem a duas curvas geradas por dois algoritmos de classificação. Como não houve intersecção entre as curvas, aquela que mais se aproxima do ponto (0,1) é a de melhor desempenho. Quando ocorre a intersecção, cada algoritmo possui uma região em que é melhor que a do outro (Faceli et al., 2011).

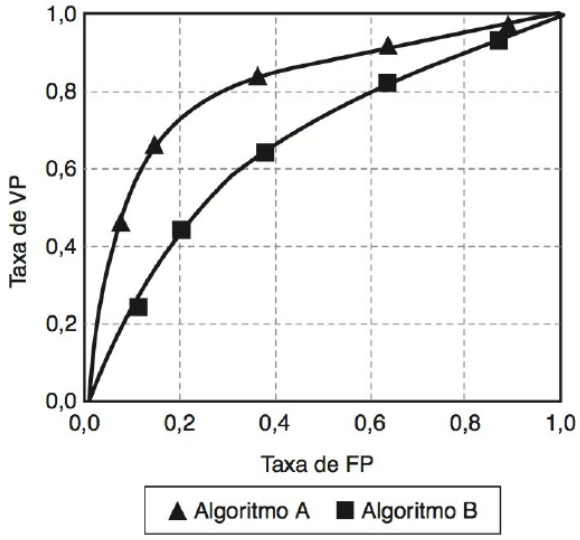

Figura 2.14: Exemplo de uma curva ROC. (Faceli et al., 2011)

A fim de comparar o desempenho dos algoritmos em termos de uma medida única por meio de um valor escalar, é extraída a área abaixo da curva ROC (AUC, do inglês *Area Under ROC Curve*). A medida AUC produz valores entre 0 e 1. Os valores mais próximos de 1 são considerados melhores(Faceli et al., 2011).

2.10 Wilcoxon signed-ranks - WSR

De acordo com Corani et al. (2017), a realização de uma comparação estatística das medidas de avaliação dos algoritmos de aprendizado de máquina é fundamental, e é normalmente realizada através de testes de significância estatística, ou simplesmente testes estatísticos. Dessa forma, os referidos testes permitem interpretar os resultados das inúmeras execuções do classificador, obtidas, por exemplo, através do *Cross-Validation*. Com isso, pode-se inferir afirmações sobre os resultados, considerando determinado classificador melhor que outro.

Os testes estatísticos podem ser separados em dois grandes grupos: paramétricos e não-paramétricos. Os testes paramétricos exigem que a distribuição dos dados experimentais apresente populações com distribuições específicas. Não apresentando normalidade na distribuição das amostras dos dados, ou caso não se tenha elementos suficientes para poder afirmar que seja uma distribuição normal, o teste estatístico utilizado deverá fazer parte do grupo de testes não-paramétricos, que são denominados de testes livres de distribuição. Isso possibilita que os testes não-paramétricos sejam aplicados em uma larga variedade de situações.

Uma alternativa entre os testes não-paramétricos, o teste de Wilcoxon signed-ranks (WSR) proposto por Wilcoxon (1945), pode ser utilizado para avaliar a performance de dois algoritmos de classificação, comparando-os por meio da diferenças negativas e positivas da medida de avaliação utilizada. As diferenças são classificadas de acordo com seus valores absolutos.

Conforme Triola (2018), o teste de WSR busca avaliar se:

- *H0*: A mediana das diferenças entres os pares combinados é igual a zero.
- *H1*: A mediana das diferenças entres os pares combinados é diferente de zero.

Dada duas medidas de avaliação pode-se seguir o procedimento descrito por Triola (2018) para a realização do WSR.

1. Para cada par de medidas, calcule a diferença d subtraindo o segundo valor do primeiro valor. Descarte quaisquer pares que tenham uma diferença de 0.
2. Ignore os sinais das diferenças e, classifique-as da menor para a maior e substitua as diferenças pelo valor da classificação correspondente. Quando as correspondências tiverem o mesmo valor numérico, associe a elas a média dos postos envolvidos no empate.
3. Atribua a cada posto o sinal da diferença de onde veio. Ou seja, insira os sinais que foram ignorados na Etapa 2.

4. Encontre a soma dos postos que são positivos. Encontre também o valor absoluto de a soma dos postos negativos.
5. Considere T a menor das duas somas encontradas no Passo 4.
6. Seja n o número de pares de dados para os quais a diferença d não é 0.
7. Determine a estatística do teste e os valores críticos com base no tamanho amostral.
   - Se $n <= 30$, a estatística do teste é T.
   - Se $n > 30$, a estatística do teste é:

$$z = \frac{T - \frac{n(n-1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \tag{2.13}$$

8. Rejeite a hipótese nula se a estatística de teste é menor ou igual ao(s) valor(es) crítico(s). Caso contrário, deixe de rejeitar a hipótese nula.

2.11 Trabalhos Relacionados

Os escorregamentos de encostas constituem um dos principais fenômenos causadores de desastres naturais, provocando, todos os anos, inúmeros prejuízos materiais e fazendo um grande número de vítimas fatais. Neste cenário, inúmeros estudos têm buscado explorar técnicas de mineração de dados ou aprendizado de máquina na tentativa de prever escorregamentos futuros a partir de dados de eventos anteriores.

Em Souza e Ebecken (2012), foi desenvolvida uma metodologia para prever ocorrências de escorregamentos induzidos por chuvas na cidade do Rio de Janeiro. Os dados de escorregamentos foram obtidos a partir de relatórios técnicos de campo e 30 pluviômetros. Além disso, foram utilizados dados referentes aos parâmetros do solo, totalizando 46 variáveis. O inventário de deslizamentos continha 1.033 amostras que foram categorizados por bairro, pois não possuíam coordenadas geográficas. Foram geradas 233 amostras de não escorregamento, levando em consideração um momento anterior ao evento do escorregamento. Os índices pluviométricos acumulados associados a cada deslizamento foram calculados de acordo com a proximidade geográfica de sua ocorrência e considerando diferentes níveis de chuva nos 6 dias anteriores. Foram desenvolvidos 3 modelos, envolvendo três abordagens: (1) previsão de deslizamentos; (2) extração de regras de associação; e (3) previsão de chuvas. A primeira abordagem utilizou RNA e Regras de Classificação para extrair um conjunto de regras interessantes com base em dois índices: suporte e confiança. A segunda abordagem utilizou o algoritmo *Apriori* para extrair regras de associação da base de dados. Por fim, a terceira abordagem gerou um modelo para prever precipitação pluviométrica. Dessa forma, tentou-se realizar a previsão de escorregamentos associando às regras com a precipitação prevista. O percentual de amostras usadas para

treinamento, teste e a validação foram de 80, 10 e 10%, respectivamente. A RNA obteve uma taxa de acerto de 94,1% para não ocorrência, 80,7% para situação em que a chuva causaria um estado de pânico e 72,4% para escorregamentos.

Em Farahmand e Aghakouchak (2013), foi desenvolvido um modelo de previsão de deslizamentos de terra usando dados de precipitação por satélite, mapas de uso do solo e informações topográficas. Para isso, eles utilizaram o algoritmo de aprendizado de máquina *Support Vector Machines* (SVM). Neste modelo foram utilizados dados de precipitação pluviométricas, dados de solo e um inventário contendo 581 eventos de escorregamentos. Foram geradas 5.810 amostras de não deslizamento de forma aleatória a partir de áreas de precipitação e de diferentes condições de uso da terra de encostas de todo o mundo. Para o treinamento e validação, foi realizada uma subamostragem aleatória com 100 repetições, sendo os dados divididos em 70% para treinamento e 30% para validação. O erro médio de 100 iterações de previsão de deslizamento foi estimado em aproximadamente 7%, sendo aproximadamente 2% de deslizamentos falsos, e aproximadamente 7% de eventos de deslizamentos classificados como não deslizamentos.

Korup e Stolle (2014), analisaram a crescente utilização de técnicas relacionadas à mineração de dados e aprendizado de máquina aplicadas na predição de escorregamento de encostas. Ressaltaram que, apesar do número considerável de pesquisas aplicadas na área, a previsão de escorregamentos no tempo e no espaço continua um desafio. Ao realizar o levantamento e avaliação sobre as pesquisas existentes, os autores chegaram à conclusão de que a maior parte dos estudos nesta área envolve a espacialização da susceptibilidade a escorregamentos, e não a previsão ou a mensuração de riscos futuros, propriamente ditos. Por fim, afirmaram que as altas taxas de sucesso precisam ser interpretadas com muito cuidado, a fim de não prejudicar a observação de que a previsão confiável de escorregamentos de encostas no espaço e no tempo requer esforços substanciais de pesquisas futuras, visto que, apesar dos avanços, vários desafios permanecem, sendo os principais: a previsão regional de escorregamentos e a previsão temporal de ruptura de taludes.

Em Tien Bui et al. (2016), foram analisados os desempenhos de 5 modelos para construção de mapas de susceptibilidade a escorregamentos. Foram avaliados SVM, RNA do tipo *Multi-Layer Perceptron* (MLP) e *Radial Basis Function* (RBF), *Kernel Logistic Regression* (KLR) e *Logistic Model Trees* (LMT). Utilizaram diversas medidas de avaliação para medir a performance dos modelos, que foram comparados a partir da AUC por meio dos testes estatísticos de Friedman e Wilcoxon. Para validar o modelo foi utilizada a técnica de *Cross-Validation*. Um total de 12 fatores condicionantes de escorregamentos foram utilizados. Os dados foram divididos aleatoriamente em uma proporção de 70:30 para treinamento e validação dos modelos, respectivamente. Foram utilizados um total de 98 deslizamentos de terra rotacionais rasos e realizaram amostragem aleatória para geração de não ocorrência, ou seja, as amostras foram geradas a partir de pontos aleatórios de locais seguros quanto a ocorrência de deslizamentos. Dentre os 5 modelos analisados, a RNA do tipo MLP obteve o melhor desempenho com AUC = 0,902.

Can et al. (2017), utilizaram RNA para produzir mapas de suscetibilidade a deslizamentos de terra. Um inventário de 196 escorregamentos foram mapeados, sendo que as amostras de não ocorrência foram criadas aleatoriamente a partir de locais sem risco de escorregamentos. Foram utilizados um total de 6 parâmetros para o desenvolvimento do modelo (litologia, declividade, altitude, índice de umidade, aspecto e índice de Vegetação por Diferença Normalizada). Os dados foram divididos em 68% para treinamento e 32% para teste. A RNA obteve uma AUC = 0,817.

Logar et al. (2017), analisaram a capacidade de Redes Neurais Artificiais para prever o deslocamento de terra induzido por chuva de dois locais distintos e de natureza muito diferente. O fluxo de terra na Eslovênia moveu-se com uma taxa de 100 mm/dia, enquanto o movimento de massa no Reino Unido é muito lento (5mm/ano em média). Os autores chegaram à conclusão que as RNA são capazes modelar com precisão movimentos específicos de movimentos de massa se uma série adequadamente longa de dados confiáveis sobre precipitação e deslocamentos estiverem disponíveis.

Pham et al. (2018b), propuseram um uma abordagem híbrida de aprendizado de máquina de *Random Subspace* (RSS) e *Classification And Regression Trees* (CART), denominado RSSCART, para previsão espacial de deslizamentos de terra. O desempenho do modelo foi avaliado por meio da AUC e o teste estatístico do Qui-quadrado foi aplicado para comparar os resultados com o desempenho de outros modelos como SVM, *Single CART* (SC), *Naive Bayes Trees* (NBT) e *Logistic Regression* (LR). Foram considerados os seguintes atributos: declividade, altitude, litologia, distância a falhas, distância a rios, distância a estradas, aspecto do talude, curvatura e precipitação pluviométrica. Foram identificados na área de estudo um total de 95 cicatrizes de deslizamento de terra. As amostras de não escorregamentos foram geradas a partir locais seguros, e a base de dados foi separada na proporção de 75% para treinamento e 25% para validação. Os autores concluíram que o desempenho do modelo RSSCART (AUC = 0,841) foi melhor em comparação com outros modelos.

Em Tehrani et al. (2019), foram implementados os algoritmos LR e AD para classificar eventos de escorregamentos e não escorregamentos. Os autores alcançaram uma AUC superior a 0,88, e utilizaram uma base de dados global da NASA com 4.542 deslizamentos de terra. Contudo, relataram que a localização dos eventos apresentavam grandes incertezas ao atribuir as coordenadas geográficas. As amostras de não escorregamentos foram geradas artificialmente, aplicando uma fração aleatória (menos de 0,5) à precipitação pluviométrica e as características topográficas de casos de escorregamentos. Foi considerado o acumulado de precipitação pluviométrica de até 11 dias anterior ao escorregamento. Ao todo foram criadas 4.542 amostras de não escorregamentos, desta forma mantendo o equilíbrio entre as classes. Utilizando a técnica de *Hold-out*, dividiram a base de dados em treinamento (67%) e teste (33%). Como atributos de entrada, utilizaram a precipitação pluviométrica, altitude, declividade, índice de vegetação por diferença normalizada e tipo de solo.

Achour e Pourghasemi (2020), utilizaram os algoritmos RF, SVM e *Boosted Regres-*

sion Tree (BRT) para prever escorregamentos de um trecho de estrada no nordeste da Argélia, com o objetivo de desenvolver mapas de suscetibilidade. Os autores obtiveram uma área sobre a curva ROC de 0,972, e tinham como inventário apenas 28 locais de escorregamentos (1871 pixels).

Liu et al. (2021) empregaram três algoritmos RF, *Gradient Boosted Regression Tree* (GBRT) e RNA do tipo MLP para realizar modelagem espacial de deslizamentos de terra rasos perto de Kvam, na Noruega. Foram utilizados um total de 86 deslizamentos de terra como inventário, sendo 10.197 pontos de localização considerados no conjunto de dados (*pixels*). Os dados foram divididos aleatoriamente na proporção de 70/30. Os autores selecionaram as amostras de não escorregamento de forma aleatória. Foram utilizados 11 fatores controladores de escorregamento. Os fatores de controle dizem respeito à geomorfologia, geologia, geoambiente e efeitos antropogênicos: declividade, aspecto, curvatura do plano, curvatura do perfil, acúmulo de fluxo, direção do fluxo, distância aos rios, teor de água, saturação, precipitação e distância das estradas. O treinamento foi realizado a partir do *cross validation*. A RF apresentou os melhores resultados com AUC = 0,990.

A Tabela 2.7 apresenta alguns resultados de pesquisas que utilizam a mineração de dados e aprendizado de máquina na predição de deslizamento de terra. Em destaque encontram-se os trabalhos que mais se enquadram com linha de pesquisa proposta nesse trabalho.

A revisão de literatura mostra que os métodos de aprendizado de máquina e mineração de dados têm sido amplamente utilizados na prevenção de escorregamentos de encostas e podem atingir um desempenho satisfatório. Contudo, a maior parte dos estudos nesta área compreendem a modelagem espacial da susceptibilidade de deslizamentos de terra e não a previsão futura propriamente dita de escorregamentos. Além disso, quase sempre, esses modelos têm sido construídos sobre pequenos inventários de escorregamentos em virtude da indisponibilidade de dados integrados e em larga escala.

Basicamente foram encontrados três trabalhos que propuseram a previsão espacial e temporal de deslizamentos de terra. Esses estudos mostraram variações nas comparações entre as medições, bem como na aplicação de técnicas de aprendizado de máquina. Além disso, os trabalhos de Souza e Ebecken (2012) e Farahmand e Aghakouchak (2013) esbarraram na disponibilidade e qualidade dos dados. Utilizaram 1.033 e 581 amostras de deslizamento de terra, respectivamente. Além disso, mencionam a baixa qualidade da geolocalização das ocorrências. No caso de Souza e Ebecken (2012), tinham disponível apenas localização do bairro onde ocorreu o evento, enquanto Farahmand e Aghakouchak (2013), citam que imprecisão da localização dos eventos variavam em um intervalo de 1 a 5km.

Os autores também não conseguiram contornar o problema do desequilíbrio de classes. A base de dados de Souza e Ebecken (2012), continha, aproximadamente, 18,40% de amostras negativas e 81,60% positivas, enquanto a base de Farahmand e Aghakouchak (2013), apresentava, aproximadamente, 9% de amostras positivas

e 91% de amostras negativas. Apesar desse desequilíbrio, não utilizaram nenhuma estratégia de amostragem, o que pode gerar uma configuração tendenciosa, devido à grande desproporcionalidade entre os escorregamentos e os não escorregamentos, tornando fácil para o modelo fornecer a saída correta, e com isso maximizando as medidas de desempenho. Ainda, esses trabalhos não mencionaram a utilização de técnicas de validação como o *Cross-Validation*, que permite estimar com maior rigor a capacidade de generalização dos modelos.

Já o trabalho de Tehrani et al. (2019) propôs um modelo de previsão de escorregamentos de encostas em escala global, sendo uma abordagem de difícil aplicação prática. Ademais, as amostras de não escorregamentos foram criadas artificialmente a partir da redução de fatores condicionantes como índices pluviométricos e características topográficas. Essa abordagem pode impor um viés aos conjuntos de dados.

Nesse cenário, é perceptível que essa área ainda encontra-se em desenvolvimento, com vários desafios e limitações a serem superadas, o que motiva este trabalho, seja para construir uma base de dados integrada, aperfeiçoar aplicações de técnicas de mineração de dados, melhorar o pré-processamento de dados analisar diferentes formas de geração de amostras de não escorregamentos e seu impacto na predição, avaliar o número de dias de precipitação anterior ao escorregamento e sua implicação na capacidade preditiva, implementar novos algoritmos, como o LGBM, na aplicação da previsão de escorregamentos de encostas e por fim contribuir tecnicamente com uma futura ferramenta capaz de auxiliar no processo de tomada de decisão, no monitoramento e na redução dos danos causados pelos escorregamentos.

Tabela 2.7: Resumo de pesquisas envolvendo predição de deslizamentos de terra com mineração de dados.

Autor e data	**Amostras***	**Algoritmos**	**Resultados**	**Finalidade**
Souza e Ebecken (2012)	**1033**	**RNA**		**Predição temporal e espacial**
Farahmand e Aghakouchak (2013)	**581**	**SVM**	**ERRO = 7%**	**Predição temporal e espacial**
Korup e Stolle (2014)	-	-	-	Levantamento de pesquisas na área
Tien Bui et al. (2016)	340	SVM	AUC = 0,900	Criação de mapas de susceptibilidade
Chakraborty e Goswami (2017)	100	RNA	R=0,98 e RMSE = 0,06	Prever fator de segurança
Can et al. (2017)	196	RNA	AUC = 0,817	Criação de mapas de susceptibilidade
Logar et al. (2017)	2	RNA	-	Previsão de deslocamento de deslizamento
Zhu et al. (2017)	2	LSSVM e GA	-	Previsão de deslocamento de deslizamento
Pham et al. (2018a)	1.295	AODE, SVM, RNA, LR e NB	AUC = 0,968	Criação de mapas de susceptibilidade
Pham et al. (2018b)	95	RSSCART, SVM, NB, RL	AUC = 0,841	Criação de mapas de susceptibilidade
Qi e Tang (2018)	148	RF, RNA, AD, RL e GBM	AUC = 0,967	Previsão de estabilidade de taludes
Xiao et al. (2019)	665	RF	AUC = 0,834	Criação de mapas de susceptibilidade
Dou et al. (2019)	8.459	PLFR, InV, CF, RNA e SVM	AUC = 0,87	Criação de mapas de susceptibilidade
Yi et al. (2019)	917	RL, NB e SVM	AUC = 0,818	Criação de mapas de susceptibilidade
Tehrani et al. (2019)	**4.542**	**RL**	**AUC = 0,88**	**Predição temporal e espacial**
Achour e Pourghasemi (2020)	28	RF, SVM e BRT	AUC = 0,972	Criação de mapas de susceptibilidade
Hanifinia et al. (2021)	95	GLM, ME, SVM e RNA	AUC = 0,879	Criação de mapas de susceptibilidade
Liu et al. (2021)	86	RF, GBRT e RNA	AUC = 0,990	Criação de mapas de susceptibilidade
Lucchese et al. (2021b)	86	RNA, FIS	AUC = 0,941	Criação de mapas de susceptibilidade

Capítulo 3

Metodologia

Como observado na seção 2.11, a maior parte dos trabalhos relacionados à predição de escorregamentos de encostas envolve a espacialização da susceptibilidade a estes fenômenos que são desenvolvidos, na maioria das vezes, sobre um pequeno conjunto de inventários de escorregamentos devido à escassez de dados integrados e em larga escala. Além disso, existe uma diversidade nos resultados da predição de escorregamentos, que pode estar relacionada, dentre outros fatores, com o tamanho do conjunto de dados utilizado e com a metodologia empregada na geração das amostras de não ocorrências de escorregamentos.

Dessa forma, um estudo de construção de uma base de dados integrada, de otimização das técnicas de mineração, do pré-processamento dos dados, da análise de diferentes formas de gerar registros de não ocorrências e seu impacto na predição, da avaliação do número de dias de precipitação anterior ao escorregamento e sua implicação na capacidade preditiva e a inserção de dados geotécnicos no modelo preditivo, poderão se mostrar relevantes para a melhoria da predição de escorregamentos, e consequentemente, fornecer uma ferramenta capaz de auxiliar o processo de tomada de decisão, o monitoramento e prevenção de danos causados por esses eventos.

Neste contexto, este estudo busca construir um *dataset* e aplicar modelos de mineração de dados que possam promover uma melhor predição de escorregamentos, e para isso serão comparados os resultados de quatro classificadores.

Neste capítulo, serão apresentados, em detalhes, todos os processos utilizados para alcançar os resultados dos modelos preditivos. A Figura 3.1 apresenta de maneira simplificada as etapas metodológicas desenvolvidas neste estudo.

A primeira etapa consiste na aquisição dos dados necessários para realizar um estudo desse tipo. Esses dados foram adquiridos por meio de fontes de dados distintas, devido à inexistência de uma base de dados que possuísse todas as informações necessárias para o desenvolvimento deste trabalho.

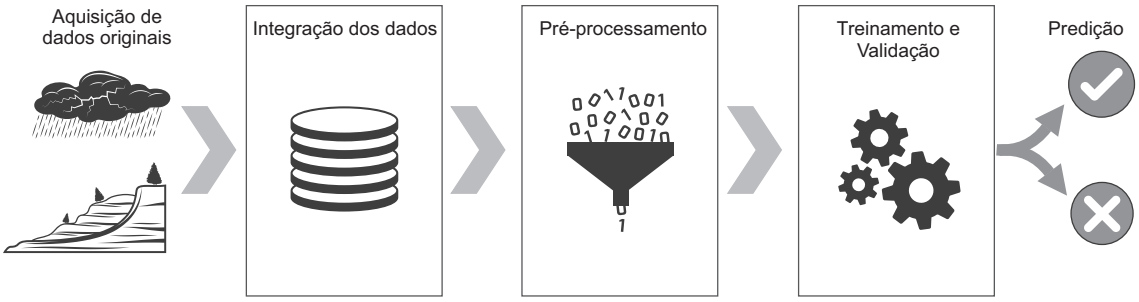

Figura 3.1: Etapas metodológicas da pesquisa

De posse dos conjuntos de dados, realizou-se a identificação e o estudo das informações disponíveis, com o objetivo de promover sua integração.

A etapa de pré-processamento foi indispensável para a adequação dos dados, sendo alicerçada em: remoção de registros incompletos e/ou duplicados, valores errados e dados inconsistentes, exclusão de registros irrelevantes ao contexto da pesquisa. Em seguida, foi necessário transformar a representação dos dados a fim de adequar os dados às necessidades dos algoritmos, sendo baseada na normalização dos dados e na adequação e padronização de formato de atributos.

Após a etapa de pré-processamento, os modelos de classificação foram construídos a partir da implementação dos algoritmos de aprendizagem de máquina. Os modelos foram avaliados, em duas etapas, por meio de cinco medidas de avaliação e os parâmetros foram otimizados de maneira que os resultados fossem maximizados.

Por fim, os resultados dos experimentos foram comparados e analisados, podendo assim, indicar os melhores algoritmos e as melhores configurações para a realização do objetivo pretendido. A seguir, cada uma das seções desse capítulo corresponde a uma das etapas desenvolvidas na investigação, detalhando os principais aspectos de cada uma.

3.1 Aquisição de dados Originais

Para obter uma base de dados, com variáveis necessárias para realizar a predição de escorregamentos de encostas, foi necessário construir uma base integrada a partir de múltiplas fontes de informação. Os itens seguintes detalham os conjuntos de dados obtidos em cada instituição.

3.1.1 Inventário de escorregamentos

Os registros de deslizamentos de terra foram cedidos pela Defesa Civil de Salvador (CODESAL). Os dados são referentes ao período de janeiro de 2004 a junho de 2021 e foram transformados e analisados utilizando o *software* livre Qgis 3.16, visto que foi

disponibilizado um arquivo georreferenciado de pontos no formato *shapefile* (SHP)[1]. Esse conjunto de dados contém 4.892 registros de escorregamentos e cada registro é constituído por: número do processo, código da vistoria, data e hora da abertura do processo, data e hora da vistoria, tipo de ocorrência, bairro, bairro prefeitura, causa, descrição e coordenadas UTM. A Tabela 3.1 ilustra um registro de escorregamento obtido dos dados da CODESAL, e a Figura 3.2 mostra a distribuição espacial dos escorregamentos registrados no município de Salvador-BA.

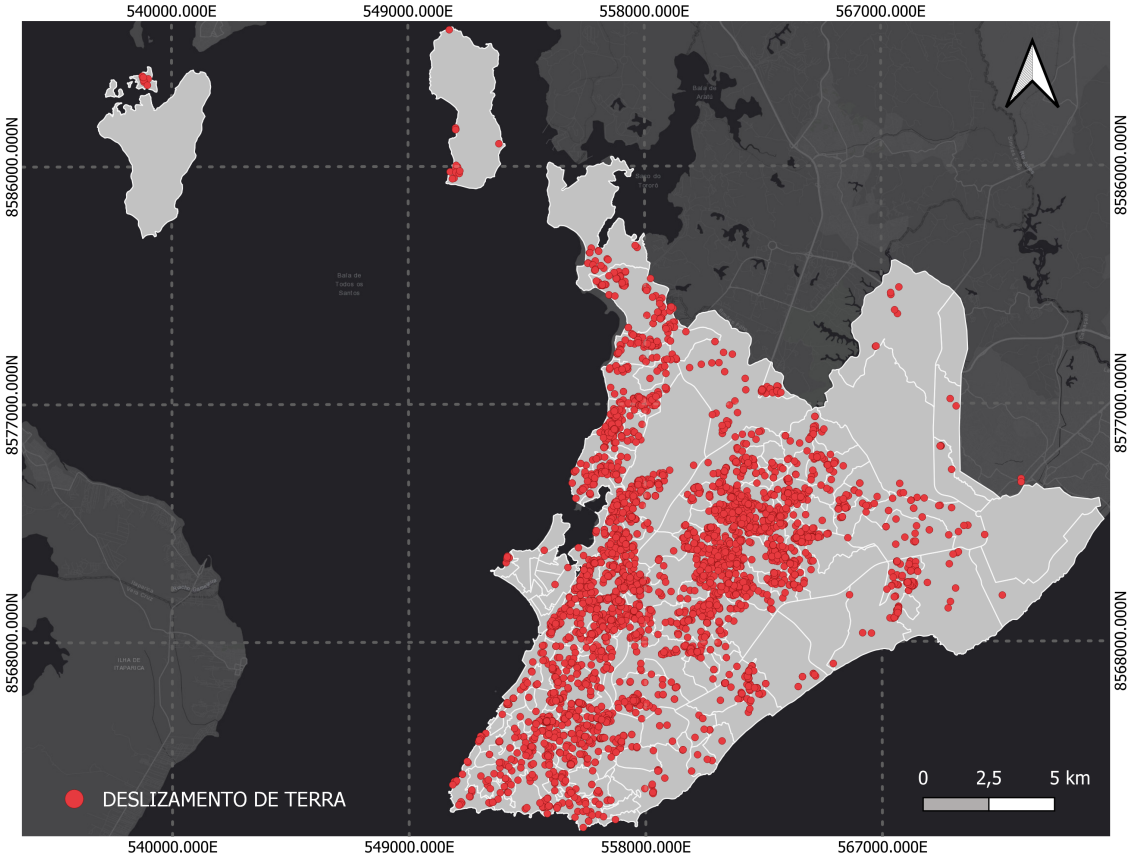

Figura 3.2: Mapa de deslizamentos de terra registrados entre janeiro de 2004 a junho de 2021 (Autor, 2022). (Sistema de referência: SIRGAS 2000/ UTM Zona 24S)

[1]SHP é um formato de arquivo para bases de dados geoespaciais e vetoriais em Sistemas de Informação Geográficas (SIG).

Tabela 3.1: Amostra de registro de escorregamento da CODESAL.

Número do Processo:	104132
Código da Vistoria:	212619
Data do Processo:	25/03/2020 15:03:01
Data da Vistoria:	27/03/2020 09:52:15
Tipo de Ocorrência:	Deslizamento de Terra
Bairro:	Liberdade
Bairro Prefeitura:	Liberdade/São Caetano
Causa:	Fortes chuvas do período, falta de contenção e ou proteção superficial da encosta.
Descrição:	Deslizamento de terra, advindo de cota superior do talude, atingindo o imóvel do requerente, causando danos materiais e queda da parede do fundo.
UTM(E):	561216,041
UTM (N):	8572616,943

3.1.2 Dados geotécnicos e geológicos

As características dos materiais que constituem o substrato das encostas são atributos importantes no processo de desencadeamento dos escorregamentos, pois cada solo possui uma resistência própria. Com isso, buscou-se dados que caracterizassem os solos dos locais onde ocorreram os escorregamentos. Esses dados foram obtidos nas seguintes instituições:

Laboratório de Geotecnia da UFBA

Os dados geotécnicos foram cedidos pelo Laboratório de Geotecnia da UFBA que disponibilizou seis arquivos em formato *raster* (TIFF), contendo dados de propriedades do solo para uma determinada região do município de Salvador. As propriedades disponíveis foram: coesão, ângulo de atrito interno e peso específico, para as condições de umidade natural (material natural) e saturada (material inundado ou saturado).

Para a construção desses dados matriciais contendo os valores de coesão, ângulo de atrito interno e peso específico do solo, foram utilizados os dados (vetoriais) geotécnicos disponíveis de amostras de blocos indeformados de solos distribuídas pela cidade de Salvador, sendo 235 amostras para a condição natural e 382 amostras para condição saturada. Para isso, foi utilizado o método de interpolação especial denominado krigagem ordinária [2] e definida a dimensão do pixel de 5 metros (Santos, 2018).

[2]Krigagem é o processo geoestatístico de estimativa de valores de variáveis distribuídas no espaço e/ou tempo, com base em valores adjacentes quando considerados interdependentes pela análise variográfica. Já a krigagem ordinária nada mais é que a sua própria versão simplificada com a média local calculada pela krigagem média (Yamamoto e Landim, 2013).

Companhia de Pesquisa de Recursos Minerais - CPRM

Foram cedidos pela CPRM os dados geológicos através de mapa georreferenciado de polígonos em formato *shapefile*. Esse mapa define as características geológicas dos limites territoriais do município de Salvador. Cada registro é composto por: domínio geológico, relevo e textura do solo. A Figura 3.3 apresenta o mapa geológico simplificado do município de Salvador.

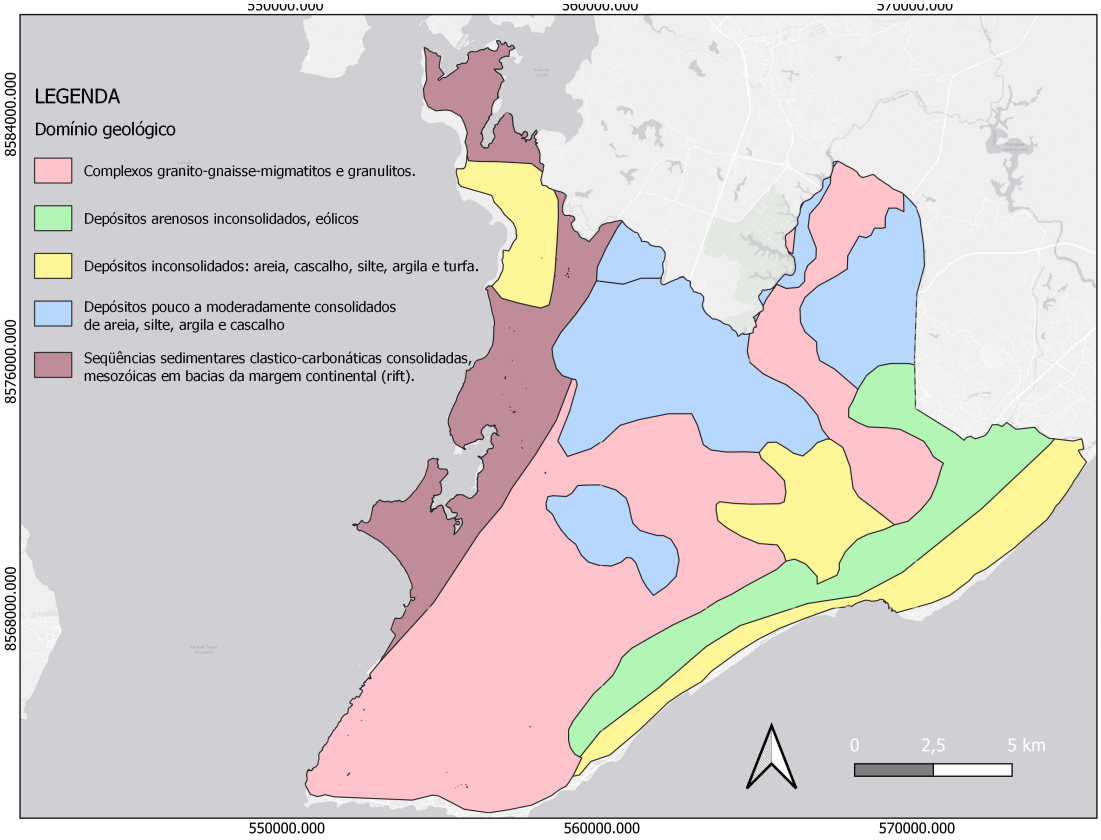

Figura 3.3: Mapa geológico simplificado município de Salvador, Bahia (Autor, 2022 - Dados base CPRM). (Sistema de referência: SIRGAS 2000/ UTM Zona 24S)

3.1.3 Dados de Precipitações Pluviométricas

Devido à grande influência da precipitação no processo de escorregamento de encostas buscou-se obter a maior quantidade possível de dados de precipitação. A Figura 3.4 exibe a distribuição espacial de todas as estações pluviométricas do município de Salvador que foram utilizadas nesse estudo. Os dados dessas estações foram obtidos das seguintes instituições:

CODESAL

Os registros de dados pluviométricos obtidos através da CODESAL são produzidos pelo Centro de Monitoramento e Alerta da Defesa Civil de Salvador (CEMADEC), composto por 31 estações pluviométricas automáticas, distribuídas de forma estratégica dentro dos limites territoriais da capital da Bahia, e são capazes de registrar, em tempo real e a cada 15 minutos, dados de precipitação pluviométrica. Foram disponibilizados dois arquivos, um com as informações das estações pluviométricas e outro com os dados de precipitação. A base de dados abrange o período entre outubro de 2016 a outubro de 2021. Os dados cedidos foram registrados em UTC (*Universal Time Coordinated*) ou GMT (*Greenwich Meridian Time*), que é a hora no Meridiano de Greenwich, na Inglaterra. As Tabelas 3.2 e 3.3 ilustram registros de dados das estações pluviométricas e da precipitação, respectivamente.

Tabela 3.2: Amostras de registros das estações da CODESAL.

Código da estação	Nome	Latitude	Longitude
28	Ilha de Maré	-12,78044447	-38,53180073
42	Retiro	-12,96530700	-38,47840100
69	Campinas de Brotas	-12,98143800	-38,47902400

Tabela 3.3: Amostras de registros de precipitação da CODESAL.

Código da estação	Data e hora	Valor (mm)
28	2016-10-08 04:45:00	0,60
28	2016-10-08 05:00:00	0,80
42	2020-06-19 03:55:00	1,40

Centro Nacional de Monitoramento e Alertas de Desastres Naturais - CEMADEN

A base de dados do CEMADEN está disponível em seu sítio eletrônico[3], sendo composta por 20 estações pluviométricas que registram automaticamente dados de precipitação a cada 10 minutos. Os dados estão disponíveis em arquivos separados por mês, a partir de setembro de 2013. Assim como os dados da CODESAL, os dados foram registrados em UTC ou GTM. A Tabela 3.4 ilustra três registros de dados de precipitação do CEMADEN.

Tabela 3.4: Amostras de registros de precipitação da CEMADEN.

ID estação	Nome	Latitude	Longitude	Data e hora	Valor(mm)
292740801A	Pirajá	-12,89870	-38,45887	2021-06-01 13:20:00	2.56
292740801A	Pirajá	-12,89870	-38,45887	2021-06-01 13:30:00	6.30
292740805A	Rio Sena	-12,88769	-38,46827	2018-01-10 10:40:00	1.37

[3]http://www2.cemaden.gov.br/mapainterativo/

Instituto Nacional de Meteorologia - INMET

A base de dados obtida do INMET é constituída de dados de três estações pluviométricas convencionais com registros diários de precipitações. O órgão que possui a estação mais antiga de Salvador, disponibilizou registros obtidos entre agosto de 1963 e outubro de 2021, conforme Tabela 3.5. A Tabela 3.6 ilustra três registros de dados diários de precipitação do INMET.

Tabela 3.5: Estações pluviométricas INMET.

ID	Nome da estação	Latitude	Longitude	Período
83229	Salvador-Ondina	-13,005278	-38,505833	1963 a 2021
A401	Salvador-Zoológico	-13,0055	-38,5058	2000 a 2021
A456	Salvador-Estação Rádio Marinha	-12,808222	-38,495944	2018 a 2021

Tabela 3.6: Amostras de registros de precipitação do INMET.

ID estação	Data	Valor(mm)
83229	2017-01-10	19.40
83229	2020-05-12	45.30
A401	2018-07-07	29.50

Instituto do Meio Ambiente e Recursos Hídricos - INEMA

A base de dados cedida pelo INEMA é composta por registros pluviométricos diários de cinco estações. Algumas destas já foram desativadas, contudo, como os registros compreendem parte do período estudado, optou-se pelo utilização desses dados. A tabela 3.7 apresenta as estações com o período de abrangência dos dados.

Tabela 3.7: Estações pluviométricas INEMA.

ID	Nome da estação	Latitude	Longitude	Período
RN-CL-01	Salvador-Itapuã	-12,931388	-38,361111	1998 a 2016
RN-CL-05	Salvador-Abaeté	-12,94469	-38,35933	2019 a 2021
RN-PR-01	Salvador-Abaeté	-12,9445407	-38,361404	2004 a 2021
RN-PR-03	Salvador-Abaeté	-12,9454	-38,361404	2012 a 2013
RN-PR-05	Salvador-Mont Serrat	-12,930959	-38,516964	2015 a 2016

Tabela 3.8: Amostras de registros de precipitação do INEMA.

ID estação	Data	Valor(mm)
RN-CL-01	2017-01-10	19.40
RN-PR-05	2015-07-09	33.20
RN-CL-05	2018-05-13	77.20

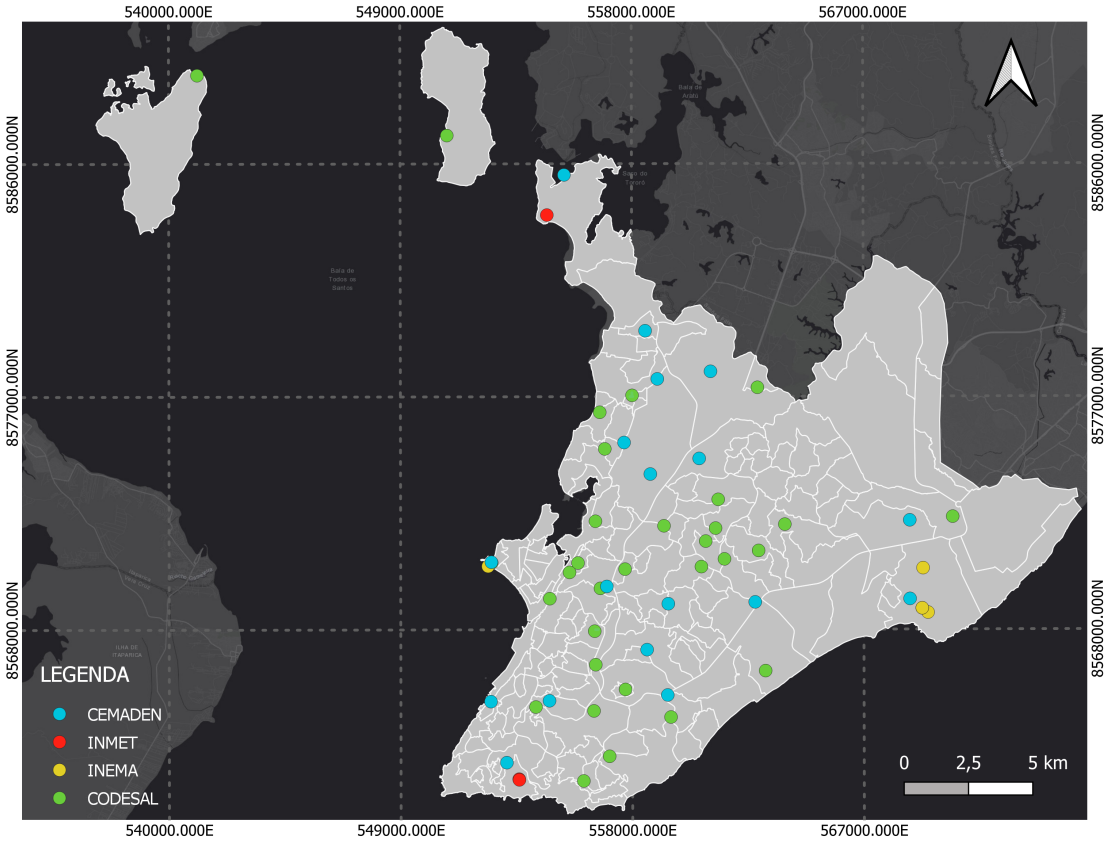

Figura 3.4: Distribuição espacial das estações pluviométricas do município de Salvador, Bahia (Autor, 2022). (Sistema de referência: SIRGAS 2000/ UTM Zona 24S)

3.1.4 Dados Geomorfométricos

Conforme visto na seção 2.3, nos estudos dos escorregamentos de encostas os dados geomorfométricos são de grande importância. Os dados geomorfométricos deste estudo foram obtidos através do banco de dados Topodata (Brasil, 2008), desenvolvido com o objetivo de oferecer informações das variáveis geomorfométricas locais básicas em cobertura nacional para utilização através do Sistema de informações Geográficas (SIG). Os dados disponibilizados estão em formato *raster* e foram processados no ano de 2011. Dentre os dados disponíveis foram utilizados o Modelo Digital de Elevação (MDE) e sua derivação de declividade. Na Figura 3.5 é exibido o mapa

hipsométrico (Altitude) do município de Salvador, enquanto a Figura 3.6 apresenta o mapa de declividade. A geomorfologia da cidade não apresenta altitudes elevadas, não ultrapassando, na região onde se concentra a cidade, altitudes de 110 m. Já no que se refere a declividade há uma predominância de relevo com declividade entre 20 e 45%, caracterizado como forte ondulado.

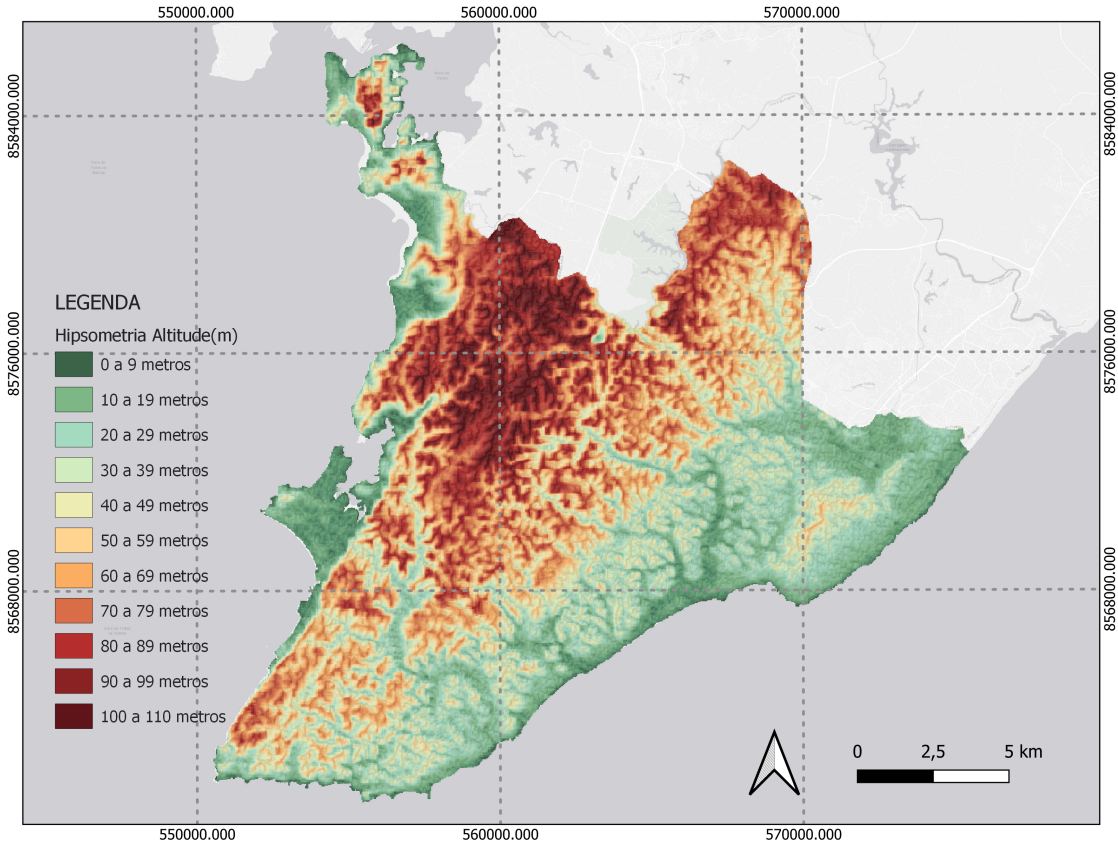

Figura 3.5: Mapa Hipsométrico do município de Salvador, Bahia (Autor, 2022 - Dados base Topodata). (Sistema de referência: SIRGAS 2000/ UTM Zona 24S)

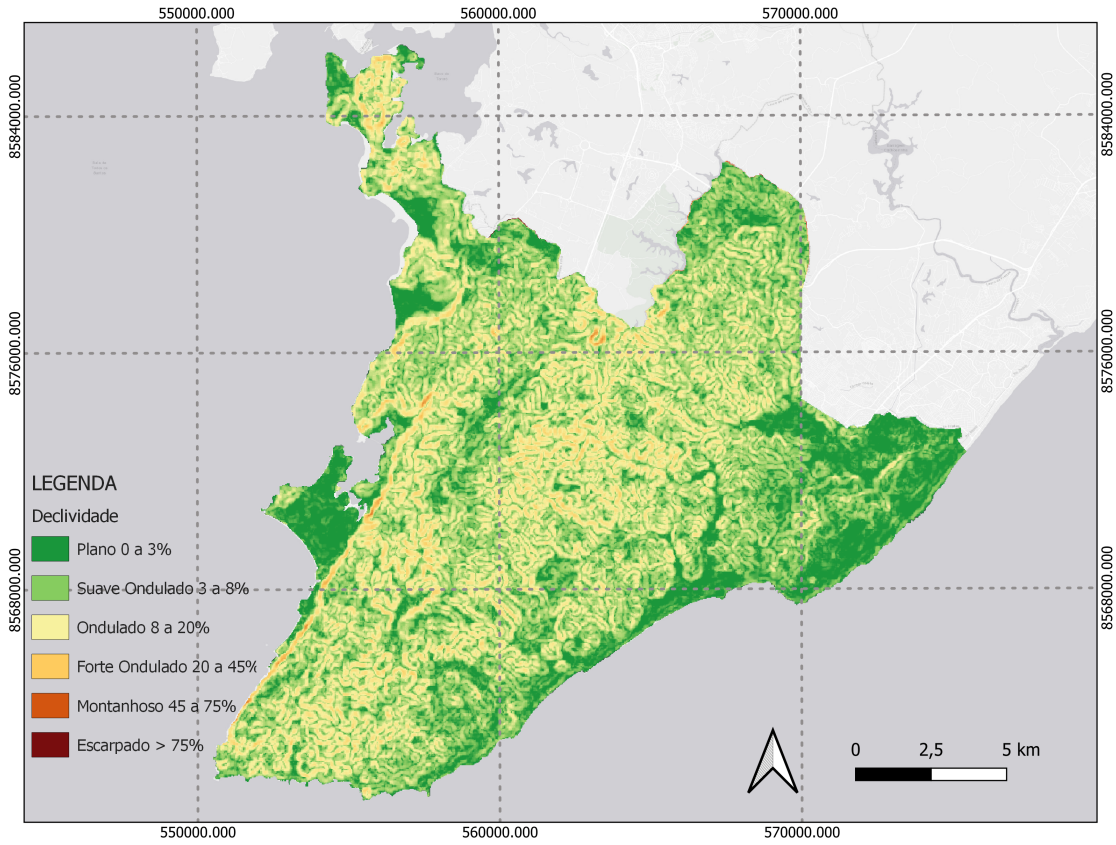

Figura 3.6: Mapa de declividade do município de Salvador, Bahia (Autor, 2022 - Dados base Topodata). (Sistema de referência: SIRGAS 2000/ UTM Zona 24S)

3.1.5 Áreas de Riscos de escorregamentos

Mapas de áreas de riscos de escorregamento foram obtidos das seguintes instituições:

CODESAL

Objetivando minimizar a ocorrência de acidentes, a CODESAL passou a atuar na prevenção. Em 2016, iniciou a elaboração do mapeamento de riscos das áreas críticas de Salvador. Foi disponibilizado um arquivo georreferenciado de polígonos no formato *shapefile*, contendo as áreas de risco mapeadas até junho de 2021.

CEMADEN e IBGE

Em 2018, o CEMADEN e o Instituto Brasileiro de Geografia e Estatística (IBGE) lançaram uma base de dados sobre a população exposta em área de desastres naturais. Estão disponíveis no sítio eletrônico do IBGE, mapas georreferenciados das áreas de riscos de desastres naturais, dentre eles, o do município de Salvador. A Fi-

gura 3.7 apresenta o mapa das áreas de risco de desastres mapeadas pela CODESAL e pelo IBGE.

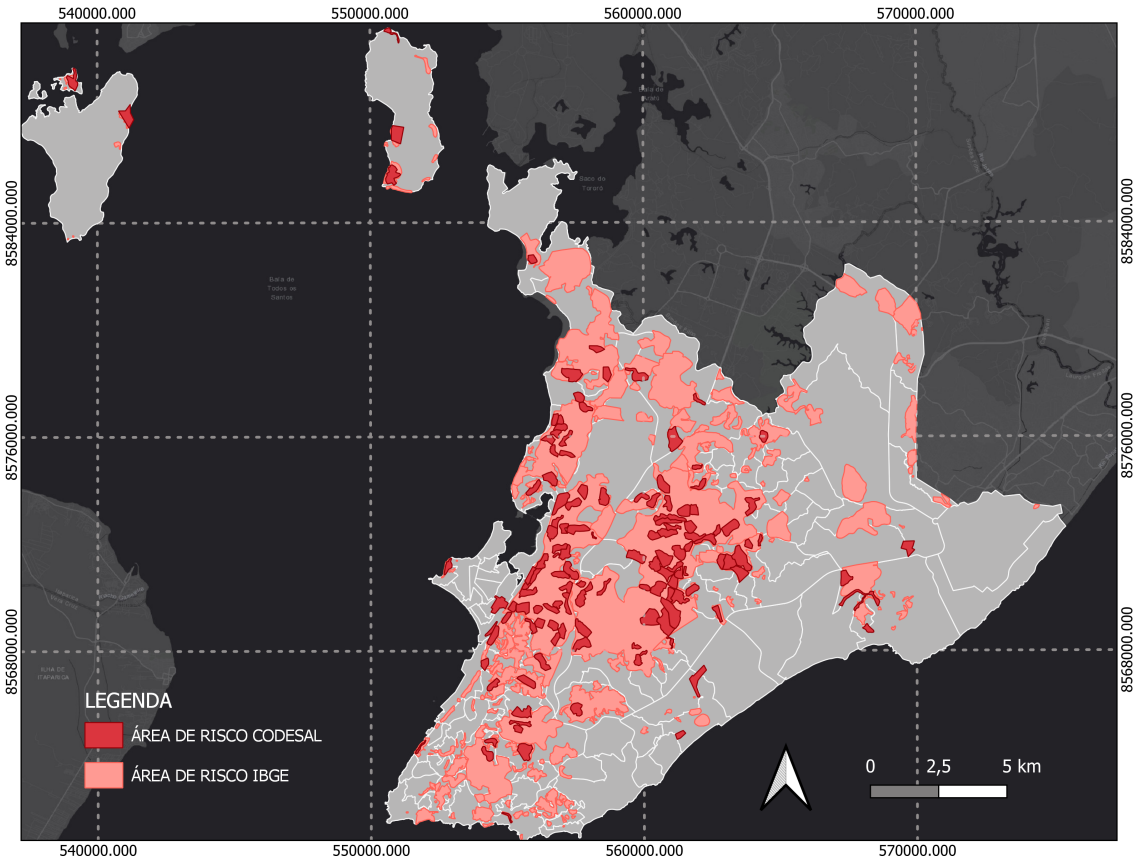

Figura 3.7: Áreas de risco de desastres do município de Salvador, Bahia (Autor, 2022 - Dados base Codesal e IBGE). (Sistema de referência: SIRGAS 2000/ UTM Zona 24S)

3.2 Integração dos dados

Para realizar a integração dos dados foram desenvolvidas diversas tarefas com o objetivo de criar uma base de dados compostas por características essenciais para aplicação de técnicas de mineração de dados e aprendizado de máquina. Boa parte dos dados adquiridos são do tipo geoespaciais, que possuem dois formatos primários: *raster*/matricial e *vector*/vetorial. A Figura 3.8 traz a representação simbólica dos arquivos do tipo matricial e vetorial. O primeiro é composto por linhas e colunas de *pixels*, em que cada *pixel* representa uma característica de uma determinada região geográfica. Já o arquivo vetorial trata-se de um tipo de dado digital que representa uma feição ou elemento gráfico, seja ela em formato de ponto, linha ou polígono, e que possui referência geoespacial.

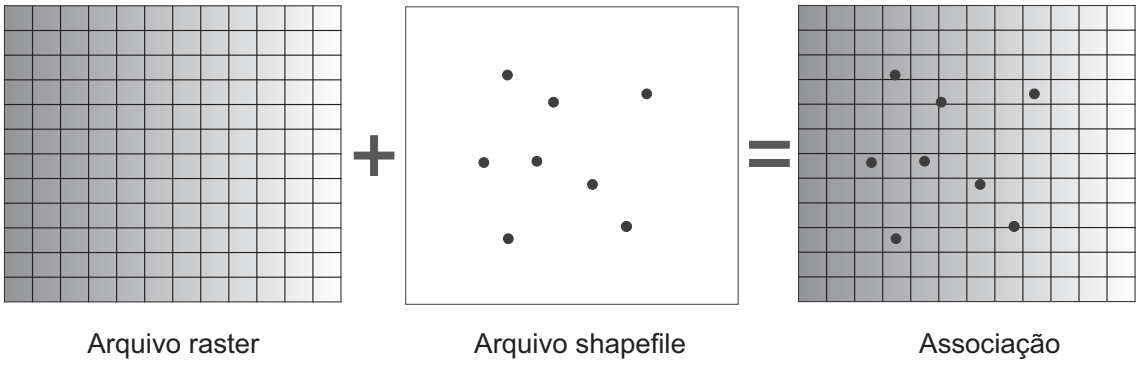

Figura 3.8: Associação de arquivos georreferenciáveis. Fonte: Autor (2021)

Para integrar os dados geoespaciais, foi necessária a utilização de um *software* de Sistema de Informação Geográfica. Para este estudo, utilizou-se o QGIS.

Desta forma, os dados de escorregamentos foram integrados com os dados geomorfométricos e geológicos por meio da localização geoespacial, visto que em um ambiente SIG ao associar dois arquivos, por meio da sua localização é possível copiar os valores dos *pixels*, onde as feições (pontos) estão sobrepostas, para a tabela de atributos do arquivo vetorial. De forma similar, esse procedimento foi realizado para integrar as ocorrências de escorregamentos com áreas de riscos.

Como os dados de precipitação pluviométrica possuem a latitude e a longitude para cada pluviômetro, foi possível determinar a distância de cada escorregamento para cada estação.

Deste modo, os dados foram integrados a partir de uma matriz de distâncias entre as estações e os eventos de deslizamentos de terra, considerando que a precipitação diária é obtida por meio da estação mais próxima que possui dados disponíveis para a data do escorregamento e período anterior.

3.3 Pré-processamento

Foram realizadas diferentes tarefas de pré-processamento fundamentais para tratar as informações existentes, de maneira que se possa evitar situações em que a predição possa ser afetada por atributos ausentes, duplicados e inconsistentes.

3.3.1 Remoção de registros

Ao realizar uma análise sobre a base de dados verificou-se a existência de registros de escorregamentos duplicados. Essa duplicidade tem origem na abertura de mais de um processo para o mesmo escorregamento. Como a abertura do processo geralmente é feita por atendimento telefônico, ocorreu que mais de uma pessoa realizou uma solicitação para o mesmo evento de escorregamento, gerando a duplicidade.

Após a exclusão dos registros duplicados, foram eliminados os registros de escorregamentos que não foram induzidos por chuvas ou que foram categorizados erroneamente como escorregamentos. Para isso, foi necessário verificar individualmente a causa e a descrição de cada ocorrência. Tratou-se de um processo exaustivo, porém necessário para que se obtivesse uma base de dados coerente.

Por fim, foram excluídos os registros que possuíam valores ausentes nos atributos de coesão, ângulo de atrito e peso específico. Esses valores ausentes têm origem nos escorregamentos que ocorreram em áreas para quais não se tinha as propriedades disponíveis nos dados coletados. Desta forma, a área de estudo foi definida em função dos locais para os quais existiam dados das propriedades geotécnicas do solo.

Ao final dessas etapas, foram contabilizados 1.864 registros descartados, restando na base de dados 3.028 registros de escorregamentos, o que corresponde a cerca de 62% da base de dados inicial.

3.3.2 Transformação dos dados

A base de dados construída possui variáveis categóricas, ou seja, uma variável nominal, sem escala, não numérica. Essas variáveis necessitam ser transformadas em valores numéricos devido a limitação dos algoritmos em conseguirem trabalhar diretamente com variáveis categóricas. Para isso, os dados serão codificados a partir de duas técnicas: *Label Encoder* e *One-Hot Encoder*. O *Label Encoder* consiste em transformar as classes categóricas em números que as representam. Já o *One-Hot Encoder* significa em transformar um atributo categórico em variáveis(colunas) binárias. Desta forma, após aplicar o *Label Encoder* é necessário transformar os números em novas colunas da base de dados com a técnica *One-Hot Encoder*, com objetivo de eliminar a magnitude dos valores que não possuem significado nesse problema.

Os atributos *domínio geológico*, *relevo* e *composição* passaram por essa transformação. As tabelas 3.9 e 3.10 exemplificam as técnicas aplicadas para o atributo *domínio geológico*.

Tabela 3.9: Codificação *Label Encoder* para o atributo *domínio geológico*.

Domínio Geológico	**Numérico**
Complexo granito-gnaisse-migmatitos e granulitos	1
Depósitos arenosos inconsolidados, eólicos, cenozoicos	2
Depósitos inconsolidados: areia, cascalho, argila e turfas	3
Depósitos pouco ou moderadamente consolidados de areia, silte, argila e cascalho	4
Sequências sedimentares clástico-carbonáticas consolidadas, mesozoicas em bacias da margem continental	5

Tabela 3.10: Codificação *One-Hot Encoder* para o atributo *domínio geológico.*

Dom_1	**Dom_2**	**Dom_3**	**Dom_4**	**Dom_5**
1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

3.3.3 Preparação dos Dados de Precipitação Pluviométrica

Conforme visto na subseção 3.1.3, o conjunto de dados de chuva é constituído por registros de precipitação pluviométrica de 4 instituições (CODESAL, CEMADEN, INMET e INEMA), e possuem intervalos que vão de 10 minutos a 24 horas, coletados em 59 pluviômetros automáticos e convencionais instalados no município de Salvador.

Como a informação de data e hora obtida por meio do inventário de escorregamentos refere-se à data de abertura do processo da ocorrência e não ao momento propriamente dito do escorregamento, optou-se por utilizar a precipitação com medições diárias (acumulado de 24 horas), com o intuito de minimizar a imprecisão da relação cronológica do momento em que ocorreu o escorregamento e os índices pluviométricos acumulados associados. Além disso, parte significativa dos escorregamentos, cerca de 25%, ocorreram em um período que a cidade de Salvador só possuía estações pluviométricas convencionais, que registram apenas precipitações diárias. Em consequência disso, os dados obtidos das estações pluviométricas com medições em intervalos de 15 e 10 minutos precisaram ser transformados em registros diários de precipitação, totalizando os dados diários de 09:01 do dia anterior às 09:00 da data do dia subsequente, conforme convenção adotada nos registros diários, sendo também necessário o ajuste dos dados para o horário local (UTC-3).

Feitas as devidas adequações, estabeleceu-se que os índices pluviométricos devem ser obtidos com dados do pluviômetro mais próximo do local onde ocorreu o escorregamento, e que possua registros de precipitação disponíveis para data do evento. Para isso, foi construída uma matriz, com a distância de cada escorregamento para cada estação pluviométrica. Por fim, foi necessário definir os índices de chuva diário e acumulado associados a cada escorregamento.

3.4 Amostras de não ocorrência e cenários de análise

Para implementar os modelos preditivos propostos, casos de não escorregamentos devem ser adicionados ao conjunto de amostras, tendo assim, duas classes possíveis: positiva (escorregamento) e negativa (não escorregamento). O emprego de amostras de não ocorrência é muito importante no processo de predição de escorregamentos,

contudo, até o momento, há pouco consenso sobre os métodos utilizados para coletar essas amostras.

Tehrani et al. (2019), criaram casos artificiais de não escorregamentos aplicando um coeficiente redutor obtido de forma aleatória (menos de 0,5) nos índices de precipitação pluviométrica e nas características topográficas a partir de registros de escorregamentos, por exemplo, se para um escorregamento X a declividade é 30°, e o coeficiente redutor é 0,4, a amostra negativa gerada a partir de X terá o valor da declividade igual a 12°. Souza e Ebecken (2012), geraram amostras de não escorregamento considerando um atraso de tempo anterior ao evento de deslizamento de terra. Por exemplo, se o deslizamento ocorreu as 22:00h, a não ocorrência terá os índices pluviométricos registrados até as 15:00h. Lucchese et al. (2021a), conceberam registros de não ocorrências de forma a avaliar a que distância da cicatriz do deslizamento deve estar a amostra de não ocorrência para que esta seja considerada segura, para isso, avaliaram três cenários distintos variando a distância até a cicatriz.

Diante do exposto, este trabalho busca avaliar o impacto dos métodos de geração de amostras de não ocorrências na capacidade de aprendizado e predição dos modelos preditivos, levando em consideração tanto o intervalo de tempo anterior ao evento de escorregamento, quanto a distância para uma encosta próxima que deslizou em momento distinto, com o objetivo de verificar a implicação de cada método na capacidade preditiva dos modelos. Com isso, não busca-se selecionar amostras facilmente classificáveis, facilitando o processo de modelagem e apresentando métricas de avaliação infladas, sem necessariamente isso correlacionar para uma boa capacidade de generalização dos modelos.

O método de geração de amostras a partir da distância para uma encosta próxima que deslizou em momento distinto é ilustrado através da Figura 3.9, onde os pontos amarelos representam escorregamentos, o ponto verde o escorregamento de referência, ou seja, a amostra positiva e o ponto laranja consiste no escorregamento mais próximo que escorregou em momento posterior. O processo de geração desse tipo de amostra é realizado calculando-se a distância entre o escorregamento de referência (ponto verde) e todos os outros escorregamentos registrados na região. Em seguida, é verificado qual desses é o escorregamento mais próximo que deslizou em uma data futura ao escorregamento referência. Com isso, as amostras são geradas a partir de locais onde a probabilidade de ocorrência de deslizamentos de terra é alta. Assim, esse tipo de amostra negativa é constituída pelas características geomorfométricas e geológicas do escorregamento mais próximo (ponto laranja) e pelas condições pluviométricas a partir da data do escorregamento de referência (ponto verde).

A Figura 3.10 ilustra como são geradas as amostras negativas a partir do próprio talude, considerando um intervalo de tempo anterior ao evento de escorregamento. São exibidas na Figura 3.10 duas encostas, a de cor verde representa a encosta que deslizou no dia 29/03/2005, ou seja, a classe positiva, enquanto a encosta de cor laranja caracteriza a amostra de não escorregamento ou classe negativa, que foi gerada

com o intervalo de um dia anterior ao deslizamento de terra. No processo de geração desse tipo de amostra tomou-se cuidado para que o intervalo de tempo anterior ao evento fosse o menor possível a fim de evitar amostras facilmente classificáveis.

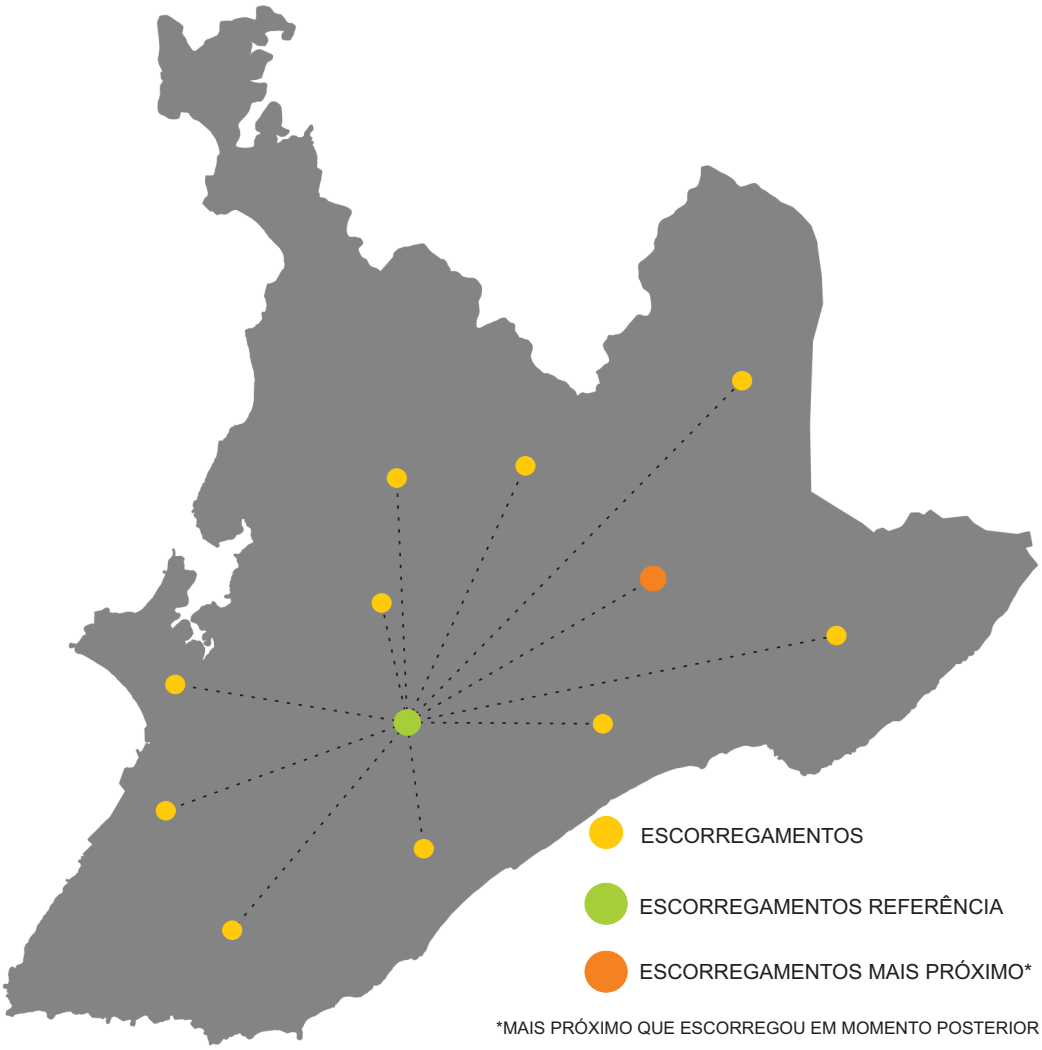

Figura 3.9: Geração de amostras negativas a partir do vizinho mais próximo que escorregou em momento posterior. Fonte: Autor (2021).

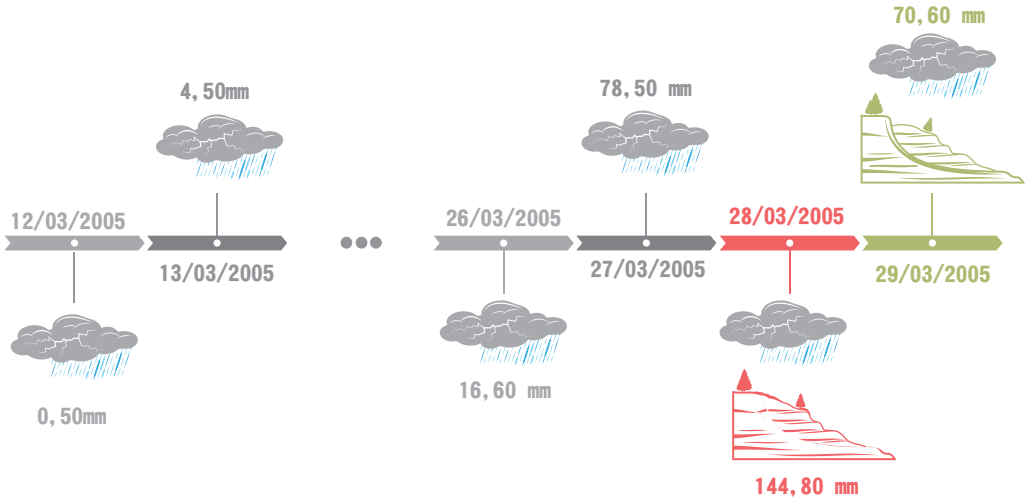

Figura 3.10: Linha do tempo com amostras positivas e negativas.

Ainda no que tange a geração de amostras negativas, sabe-se que seria possível a geração de infinitas amostras de não escorregamentos com relação temporal, contudo, isso torna o processo inviável. Desta forma, considerou-se, a princípio, a proporção de 50/50 de amostras positivas e negativas, sendo um quantitativo razoavelmente representativo para as amostras negativas, além de manter o equilíbrio das classes.

Conforme visto na Seção 2.3.1, ainda não foi possível estabelecer um intervalo de tempo ideal para caracterizar as condições pluviométricas em que são prováveis as ocorrências de escorregamentos, visto que esse intervalo pode variar conforme cada região. Desta forma, foram experimentados os valores diários e acumulados de 4, 8 e 16 dias anteriores ao evento de escorregamento para verificação. Essas valores foram definidos com base na literatura apresentada na seção 2.3.1.

Nesse contexto, dada uma determinada encosta pode-se definir uma função de predição de escorregamento como:

$$f(G, S, P) = \{0, 1\} \tag{3.1}$$

onde:
G - Propriedades Geomorfométricas da encosta;
S - Características Geológicas e propriedades Geotécnicas;
P - Precipitação diária e acumulados.

Desta forma, dado o instante do escorregamento (I_k), pode-se definir as amostras positivas de escorregamento através da Equação 3.2.

$$A_p = (G, S, P_{[I_k - N_h, I_k - D_a - (N_h - 1)]}) = 1 \tag{3.2}$$

Onde:
A_p - Amostra positiva;

G - Propriedades Geomorfométricas da encosta;
S - Características Geológicas e propriedades Geotécnicas;
P - Precipitação diária e acumulados;
I_k - Instante do escorregamento;
N_h - Número de dias de antecedência da previsão;
D_a - Número de dias de chuva acumulada anterior ao evento.

Para as amostras negativas geradas a partir das características da própria encosta, na qual é levado em consideração um intervalo de tempo anterior ao evento de escorregamento, e dado o momento do escorregamento (I_k), têm-se a Equação 3.3.

$$A_{nt} = (G, S, P_{[I_k - N_h - I_a, I_k - D_a - I_a - (N_h - 1)]}) = 0 \tag{3.3}$$

Onde:
A_{nt} - Amostra negativa;
G - Propriedades Geomorfométricas da encosta;
S - Características Geológicas e propriedades Geotécnicas;
P - Precipitação diária e acumulados;
I_k - Instante do escorregamento;
N_h - Número de dias de antecedência da previsão;
I_a - Número de dias anterior a ser considerado para a amostra negativa;
D_a - Número de dias de chuva acumulada anterior ao evento.

Para as amostras negativas geradas a partir da encosta mais próxima que escorregou em momento posterior, dado o momento do escorregamento (I_k):

$$A_{nd} = (G_v, S_v, P_{[I_k - N_h, I_k - D_a - (N_h - 1)]}) = 0 \tag{3.4}$$

Onde:
A_{nd} - Amostra negativa;
G_v - Propriedades Geomorfométricas da encosta mais próxima;
S_v - Características Geológicas e propriedades Geotécnicas da encosta mais próxima;
P - Precipitação diária e acumulado;
I_k - Instante do escorregamento;
N_h - Número de dias de antecedência da previsão;
D_a - Número de dias de chuva acumulada anterior ao evento. A partir das Equações 3.2, 3.3 e 3.4 pode-se estabelecer cenários de análise em função do tipo de amostra de não ocorrência e em função do tempo de antecedência em que se pretende realizar a predição.

Neste trabalho consideramos dezesseis cenários que são descritos a seguir:

Cenário 1 e 2 - Amostras negativas considerando um intervalo de tempo anterior ao escorregamento:

Para o cenário 1, considerou-se 1 dia de antecedência de previsão e índices pluviométricos com 16 dias anteriores ao escorregamento, tem-se: I_k = d (dia do escorre-

gamento); $N_h = 1$ dia; $I_a = 1$ dia; $D_a = 16$ dias. Substituindo nas equações 3.2 e 3.3 temos:

$$A_p = (G, S, P[_{d-1,d-16-(1-1)}]) = (G, S, P[_{d-1,d-16}])$$

$$A_{nt} = (G, S, P[_{d-1-1,d-1-16-(1-1)}]) = (G, S, P[_{d-2,d-17}])$$

A Tabela 3.11 permite uma visualização mais ampla desse cenário, pois levou-se em consideração duas situações devido da imprecisão da hora exata do escorregamento.

Tabela 3.11: Cenário - Amostras negativas com 1 dia anterior ao escorregamento e previsão com 1 dia de antecedência.

Antecedência 1 dia	**Cenário 1**	Chuva até 9h do dia do escorregamento	
		Escorregamento	**Não escorregamento**
	Propriedades Geormoformétricas	G	G
	Caract. Geológicas e Geotécnicas	S	S
	Precipitação diária	[d-1, d-16]	[d-2, d-17]
	Precipitação acumulada	[d-1, d-16]	[d-2, d-17]
	Previsão de precipitação	d*	d-1
	Cenário 2	Chuva até 9h do dia posterior ao escorr.	
		Escorregamento	**Não escorregamento**
	Propriedades Geormoformétricas	G	G
	Caract. Geológicas e Geotécnicas	S	S
	Precipitação diária	[d-1, d-16]	[d-2, d-17]
	Precipitação acumulada	[d-1, d-16]	[d-2, d-17]
	Previsão de precipitação	[d+1**, d*]	[d*, d-1]
	Previsão de precipitação acum.	d*	d-1

* Precipitação acumulada das 09:01h do dia d+1 até 09:00 do dia d;

** Precipitação acumulada das 09:01h do dia d+2 até 09:00 do dia d+1;

Para os demais cenários com o mesmo tipo de amostra negativa, realizou-se o procedimento análogo ao apresentando. Os cenários com suas respectivas referências são listados na tabela 3.12.

Tabela 3.12: Lista de cenários com amostras negativas de 1 dia anterior ao escorregamento e 16 dias de chuva acumulada.

Antecedência	**Nome**	**Referência**
1 dia	Cenário 1*	Tabela 3.11
	Cenário 2 **	Tabela 3.11
2 dias	Cenário 3*	Tabela A.1
	Cenário 4 **	Tabela A.1
3 dias	Cenário 5*	Tabela A.2
	Cenário 6 **	Tabela A.2

* Chuva até 9h do dia do escorregamento;
** Chuva até 9h do dia posterior ao escorregamento.

Cenário 7 e 8 - Amostras negativas a partir da encosta mais próxima que escorregou em momento posterior

Para um cenário considerando 1 dia de antecedência de previsão e índices pluviométricos com 16 dias anteriores ao escorregamento, tem-se: I_k = d (dia do escorregamento); $N_h = 1$ dia; $D_a = 16$ dias. Substituindo nas Equações 3.2 e 3.4 temos:

$$A_p = (G, S, P_{[d-1,d-16-(1-1)]}) = (G, S, P_{[d-1,d-16]})$$

$$A_{nd} = (G_v, S_v, P_{[d-1,d-16-(1-1)]}) = (G, S, P_{[d-1,d-16]})$$

A Tabela 3.13 apresenta de forma detalhada o cenário, considerando duas situações a partir dos dados.

Tabela 3.13: Cenário - Amostras negativas geradas a partir da encosta mais próxima que escorregou em momento posterior e previsão com 1 dia de antecedência.

	Cenário 7	Chuva até 9h do dia do escorregamento	
		Escorregamento	**Não escorregamento**
	Propriedades Geormoformétricas	G	G_v
	Caract. Geológicas e Geotécnicas	S	S_v
	Precipitação diária	[d-1, d-16]	[d-1, d-16]
	Precipitação acumulada	[d-1, d-16]	[d-1, d-16]
	Previsão de precipitação	d*	d*
Antecedência			
1 dia	**Cenário 8**	Chuva até 9h do dia posterior ao escorr.	
		Escorregamento	**Não escorregamento**
	Propriedades Geormoformétricas	G	G_v
	Caract. Geológicas e Geotécnicas	S	S_v
	Precipitação diária	[d-1, d-16]	[d-1, d-16]
	Precipitação acumulada	[d-1, d-16]	[d-1, d-16]
	Previsão de precipitação	[d+1**, d*]	[d+1**, d*]
	Previsão de precipitação acum.	d	d

* Precipitação acumulada das 09:01h do dia d+1 até 09:00 do dia d;
** Precipitação acumulada das 09:01h do dia d+2 até 09:00 do dia d+1;

Para os demais cenários com o mesmo tipo de amostra negativa, realizou-se o procedimento análogo ao apresentando. Os cenários com suas respectivas referências são listados na Tabela 3.14.

Tabela 3.14: Lista de cenários com Amostras negativas geradas a partir da encosta mais próxima que escorregou em momento posterior e 8 dias de chuva acumulada.

Antecedência	**Nome**	**Referência**
1 dia	Cenário 7*	Tabela 3.13
	Cenário 8 **	Tabela 3.13
2 dias	Cenário 9*	Tabela A.3
	Cenário 10 **	Tabela A.3
3 dias	Cenário 11*	Tabela A.4
	Cenário 12 **	Tabela A.4

* Chuva até 9h do dia do escorregamento;
** Chuva até 9h do dia posterior ao escorregamento.

Com o objetivo de avaliar quantos dias de chuva acumulada é mais efetiva na predição de escorregamentos, foram criados mais 4 cenários, sendo 2 com 8 dias de chuva acumulada e mais 2 com 4 dias de chuva acumulada. Esses cenários são listados na Tabela 3.15

Tabela 3.15: Lista de cenários gerados para 8 e 4 dias.

Dias de chuva acum.	**Antecedência**	**Nome**	**Referência**
8 dias	1 dia	Cenário 13*	Tabela A.5
		Cenário 14 **	Tabela A.5
4 dias	1 dia	Cenário 15*	Tabela A.6
		Cenário 16 **	Tabela A.6

* Amostras negativas com 1 dia anterior ao escorregamento;

** Amostras negativas geradas a partir da encosta mais próxima que escorregou em momento posterior.

Ao final desta etapa de pré-processamento de dados, foram criados dezesseis cenários de análise. Como os índices pluviométricos mudam em função de cada cenário, foram gerados dezesseis bases de dados distintas, uma para cada cenário de análise. A tabela 3.16 ilustra o conjunto de atributos da base de dados para o cenário 1. A mudança de uma base para outra ocorre em função do número dias de chuva e seus respectivos acumulados, da antecedência preditiva e da consideração do dia posterior ao evento de escorregamento.

Tabela 3.16: Atributos da base de dados.

Atributo	**Sigla**
Coesão (Kpa)	COESAO
Ângulo de atrito (°)	A_ATRITO
Peso Específico (KN/m^3)	PESO
Altitude (m)	ALT
Declividade (%)	DECL
Precipitação dia evento	PREC_DIA_EVE
Precipitação dia -1 (mm)	PREC_DIA_MEN1
Precipitação dia -2 (mm)	PREC_DIA_MEN2
Precipitação dia -3 (mm)	PREC_DIA_MEN3
Precipitação dia -4 (mm)	PREC_DIA_MEN4
Precipitação dia -5 (mm)	PREC_DIA_MEN5
Precipitação dia -6 (mm)	PREC_DIA_MEN6
Precipitação dia -7 (mm)	PREC_DIA_MEN7
Precipitação dia -8 (mm)	PREC_DIA_MEN8
Precipitação dia -9 (mm)	PREC_DIA_MEN9
Precipitação dia -10 (mm)	PREC_DIA_MEN10
Precipitação dia -11 (mm)	PREC_DIA_MEN11
Precipitação dia -12 (mm)	PREC_DIA_MEN12
Precipitação dia -13 (mm)	PREC_DIA_MEN13
Precipitação dia -14 (mm)	PREC_DIA_MEN14
Precipitação dia -15 (mm)	PREC_DIA_MEN15
Precipitação dia -16 (mm)	PREC_DIA_MEN16
Acumulado 48 (mm)	ACUM_48
Acumulado 72 (mm)	ACUM_72
Acumulado 96 (mm)	ACUM_96
Acumulado 120 (mm)	ACUM_120
Acumulado 144 (mm)	ACUM_144
Acumulado 168 (mm)	ACUM_168
Acumulado 192 (mm)	ACUM_192
Acumulado 216 (mm)	ACUM_216
Acumulado 240 (mm)	ACUM_240
Acumulado 264 (mm)	ACUM_264
Acumulado 288 (mm)	ACUM_288
Acumulado 312 (mm)	ACUM_312
Acumulado 336 (mm)	ACUM_336
Acumulado 360 (mm)	ACUM_360
Acumulado 384 (mm)	ACUM_384
Acumulado 408 (mm)	ACUM_408
Domínio Geológico-Categoria 1	DOM_01
Domínio Geológico-Categoria 2	DOM_02
Domínio Geológico-Categoria 3	DOM_03
Domínio Geológico-Categoria 4	DOM_04
Domínio Geológico-Categoria 5	DOM_05
Relevo-Categoria 1	REL_01
Relevo-Categoria 2	REL_02
Relevo-Categoria 3	REL_03
Composição-Categoria 1	COMP_01
Composição-Categoria 2	COMP_02
Composição-Categoria 3	COMP_03
Composição-Categoria 4	COMP_04
Composição-Categoria 5	COMP_05
Composição-Categoria 6	COMP_06
Área de Risco Codesal	AREA_R_CODESAL
Área de Risco IBGE	AREA_R_IBGE
Classe	CLASSE

3.5 Classificação

Para este trabalho foram escolhidos os classificadores Árvore de Decisão (AD), *Random Forest* (RF), *Light Gradient Boosting Machine* (LGBM) e Redes Neurais Artificiais (RNA) do tipo *Multi-Layer Perceptron*, para que seja avaliada a aplicação desses algoritmos na predição de escorregamentos de encostas. Esses classificadores foram implementados em linguagem de programação *Python* através da biblioteca *Scikit-learn*.

Os algoritmos utilizados apresentam distintas abordagens de classificação, são amplamente utilizados e bem documentados. Além disso, foram escolhidos dois algoritmos com abordagem convencional e dois algoritmos combinatórios (*ensemble*). Segundo Kelleher et al. (2015), os algoritmos *ensemble* têm a vantagem de incluir inúmeros classificadores como candidatos à combinação de preditores, e a garantia teórica de que a sua performance será, pelo menos, tão boa quanto a escolha ótima entre os candidatos, oferecendo, desta forma, uma alternativa flexível comparado ao métodos convencionais. Por fim, os algoritmos AD, RF e LGBM ainda não foram explorados para o problema de predição de escorregamentos no tempo e no espaço, conforme análise de trabalhos relacionados, sendo que são algoritmos apresentam bons resultados em diversos tipos de problemas.

Para analisar os resultados, utilizou-se cinco medidas de avaliação: F1-score, acurácia (ACC), *precision*, *recall* e área sob a curva ROC (AUC). A AUC trata-se de uma das medidas mais utilizadas para avaliar classificadores. Ela leva em consideração tanto os casos verdadeiros positivos (deslizamento que foi classificado como deslizamento) quanto os falsos positivos (deslizamento que foi classificado como não deslizamento). A *recall* e a *precision* também são medidas amplamente utilizadas na avaliação de modelos preditivos. A partir da *precision* pode-se contabilizar, dentre todos os deslizamentos apontados pelo modelo, qual a fração destes que são de fato deslizamentos. Já o *recall* mensura quantos casos de deslizamentos foram descobertos pelo modelo, dentre todos os casos deslizamentos presentes em nossa base de dados. O F1-score é dado pela harmônica entre a *recall* e a *precison*. Por fim, a ACC mede o desempenho global do modelo.

Os experimentos foram conduzidos conforme o fluxo de trabalho apresentado na Figura 3.11. Preliminarmente, 20% do conjunto de dados foi reservado para testes independentes adicionais e, para isso realizou-se uma amostragem aleatória estratificada. Essa técnica evita que se tenha amostras enviesadas, ou seja, com muitos registros de uma determinada classe. Por sua vez, experimentos usando o protocolo de validação cruzada com 10 *folds* foram realizados sobre os 80% restantes dos dados.

A otimização dos hiperparâmetros foi feita através da técnica *grid search*, representada pelo retângulo azul na Figura 3.11. Com base em uma grade de hiperparâmetros, o *grid search* avalia todas as combinações, com o objetivo de encontrar a que otimize o desempenho de cada classificador. Para cada combinação criada, ou

seja, para cada modelo construído é executado o método de validação cruzada com 10 *folds*, a fim de avaliar o grau e a capacidade de generalização de cada modelo. Para cada iteração da validação cruzada são computadas as medidas de avaliação. Ao término do processo, as performances estimadas de cada modelo, são utilizadas para calcular a performance média. Foi utilizada a medida F1-score (F1) para a escolha do melhor modelo. Ao final da execução do *grid search* foi escolhido o modelo que apresentou o melhor desempenho médio. Por fim, esse modelo foi aplicado sobre o conjunto dados reservado para testes adicionais, com o objetivo de verificar a capacidade preditiva desses modelos em dados inéditos adicionais, não utilizados em qualquer etapa anterior de otimização dos modelos.

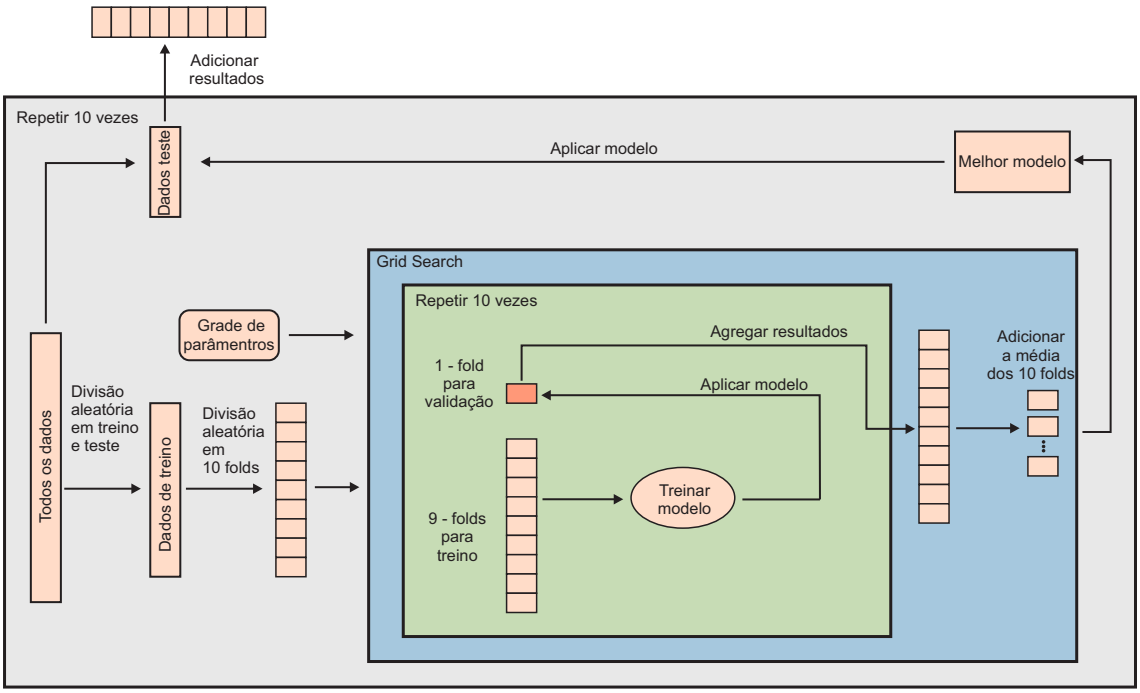

Figura 3.11: Fluxo de trabalho de otimização e classificação. Fonte: Autor (2021).

Este processo foi repetido dez vezes, sendo em cada execução atribuída uma nova semente de aleatoriedade, com o objetivo de embaralhar os dados de maneira distinta da iteração anterior. Desta forma, os conjuntos de dados de treino e teste variaram a cada iteração e em cada rodada da validação cruzada.

A Tabela 3.5 apresenta os parâmetros e os valores testados durante a execução do *grid search* para cada classificador e a tabela 3.17 exibe as sementes de aleatoriedade utilizadas durante as dez iterações.

Tabela 3.17: Sementes de aleatoriedade utilizadas em cada iteração.

iteração	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Semente	42	206	1003	435	34	59	234	890	1567	678

Tabela 3.18: Parâmetros e valores utilizados no *grid search*.

Classificador	**Parâmetros**
AD	Random state: [42*]; Critério:[‘gini’, ‘entropy’]; Profundidade máxima: [1, 2, ..., 32]; Amostras mínimas divididas: [2, 3, ..., 10]; Mínimo de amostras necessárias para estar em um nó folha: [1, 2, ..., 10].
RF	Bootstrap: [True, False]; Profundidade máxima: [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, None]; Recurso máximo: [‘auto’, ‘sqrt’]; Mínimo de amostras necessárias para estar em um nó folha: [1, 2, 4]; Mínimo de amostras para dividir um nó interno: [2, 5, 10]; Número de árvores na floresta: [200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000]; Random state:[42*].
LGBM	Taxa de aprendizado: [0.005, 0.01]; Número de árvores na floresta: [300, 600, 1000, 1200, 1500, 2000]; Máximo de folhas de árvore: [6,8,12,16]; Boosting type: [‘gbdt’]; Objetivo: [‘binary’]; Random state: [42*]; Proporção de subamostra de colunas ao construir cada árvore: [0.64, 0.65, 0.66]; subamostra: [0.7,0.75]; Reg alpha: [1,1.2]; Reg lambda: [1,1.2,1.4].
RNA	Ativação: [‘relu’, ‘tanh’]; Tamanho das camadas ocultas: [(5,), (25,), (50,), (75,), (100,), (125,), (150,), (175,), (200,), (250,), (300,), (325,), (350,)]; Solucionador: [‘adam’, ‘lbfgs’]; Taxa de aprendizado: [‘constant’, ‘adaptive’]; Alfa: [0,1, 0,01,..., 0,000000001]; Número máximo de iterações = 10000; Random state: [42*].

* Valor utilizado na primeira iteração.

3.5.1 Testes Estatísticos

Para validar a comparação entre os classificadores foi aplicado o teste *Wilcoxon signed-rank* - WSR (Wilcoxon, 1945). Para isso, foi utilizado o *software* Microsoft Office Excel com o suplemento *Real Statistics*, com grau de confiança de 95%.

Capítulo 4

Resultados

Neste capítulo serão apresentados e discutidos os resultados obtidos pelo estudo e aplicação da metodologia descrita no Capítulo 3, com o objetivo de apresentar os resultados obtidos com o emprego dos modelos preditivos para diversos cenários de ocorrência de deslizamento de terra.

4.1 Classificação

Nessa seção serão apresentados os valores das medidas de avaliação obtidos pelo classificadores, *Random Forest* (RF), Redes Neurais Artificiais (RNA) do tipo *Multi-Layer Perceptron*, Árvore de Decisão (AD) e *Light Boosting Gradient Machine* (LBGM). Foram obtidos os valores médios de F1, ACC, AUC, *Recall* e *Precision*, para cada cenário de análise. A Tabela 4.1 apresenta um resumo dos cenários analisados.

4.1.1 Resultados

Todos os cenários desta subseção foram gerados considerando dezesseis dias de chuvas e seus respectivos acumulados. Os resultados estão apresentados nas tabelas 4.2, 4.3, 4.4, 4.5, 4.6 e 4.7. A medida alvo utilizada para otimização e seleção dos modelos durante o aprendizado foi a *F1-score*, portanto, o modelo com maior *F1-score*, para cada classificador, foi o escolhido. Foram comparados estatisticamente os dois modelos que apresentaram maior AUC em cada cenário, com o objetivo de verificar se há diferenças significantes entre os resultados e, para isso, foi utilizado o teste não paramétrico de *Wilcoxon-signed-rank*.

A avaliação de performance dos modelos selecionados evidenciou um desempenho satisfatório, com *F1-score* superior a $0,810 \pm 0,005$ e AUC superior a $0,809 \pm 0,005$, para todos os modelos. Os algoritmos LGBM e RF apresentaram resultados ainda melhores, obtendo valores médios de *F1-score* superiores a $0,929 \pm 0,002$ e AUC superiores a $0,930 \pm 0,002$.

Tabela 4.1: Resumo dos cenários analisados.

Nome	Amostras Negativas	Antecedência	N° de dias de chuva Acum.
Cenário 1	Tipo 1	1 dia	16 dias
Cenário 2*	Tipo 1	1 dia	16 dias
Cenário 3	Tipo 1	2 dias	16 dias
Cenário 4*	Tipo 1	2 dias	16 dias
Cenário 5	Tipo 1	3 dias	16 dias
Cenário 6*	Tipo 1	3 dias	16 dias
Cenário 7	Tipo 2	1 dia	16 dias
Cenário 8*	Tipo 2	1 dia	16 dias
Cenário 9	Tipo 2	2 dias	16 dias
Cenário 10*	Tipo 2	2 dias	16 dias
Cenário 11	Tipo 2	3 dias	16 dias
Cenário 12*	Tipo 2	3 dias	16 dias
Cenário 13	Tipo 1	1 dia	8 dias
Cenário 14	Tipo 2	1 dia	8 dias
Cenário 15	Tipo 1	1 dia	4 dias
Cenário 16	Tipo 2	1 dia	4 dias

Tipo 1: Amostras negativas geradas a partir do próprio talude com um dia anterior ao evento de deslizamento de terra.
Tipo 2: Amostras negativas geradas a partir da encosta mais próxima que deslizou em momento posterior.
* Cenários em que foi considerado chuva até 9h do dia posterior ao escorregamento.

Com exceção da *precision*, o LGBM obteve, significativamente, os melhores resultados em cada um dos cenários (Wilcoxon signed-rank, $\alpha < 0{,}05$). Já o RF apresentou significativamente (Wilcoxon signed-rank, $\alpha < 0{,}05$), uma melhor *precision* em quase todos os cenários.

Como os falsos negativos (escorregamento classificado como não escorregamento) são mais críticos que os falsos positivos (não escorregamento classificado como escorregamento) admiti-se que o modelo aponte um maior número de falsos positivos em busca de estimar o maior número possível de escorregamentos e, consequentemente, espera-se uma *precision* menor e uma maior *recall*. Desta forma, por ter apresentado uma melhor *recall*, o LGBM é o algoritmo capaz de identificar corretamente a maioria dos casos de interesse (escorregamentos), e ainda apresenta uma boa capacidade de identificar de forma correta a classe negativa (não escorregamentos). Observou-se, também, que o algoritmo apresentou um bom equilíbrio entre a *recall* e a *precision* e um baixo desvio padrão em todas as medidas, o que dá consistência e denota uma boa confiabilidade do modelo.

Houve uma diferença de desempenho dos modelos em função do método de geração de amostras negativas. As bases de dados com amostras geradas a partir do vizinho

mais próximo permitiram maior acréscimo na performance dos classificadores, atingindo resultados ainda mais expressivos. Esse acréscimo é mais notório na RNA. Além disso, constata-se que o algoritmo LGBM aumentou a sua *precision* superando em alguns casos a RF.

As RNAs não conseguiram apresentar o mesmo desempenho dos demais classificadores, obtendo os menores valores de medidas de avaliação em todos os cenários estudados. Além disso, de maneira geral, apresentaram as maiores variações em relação às médias.

Ao analisar o desempenho da AD nas bases com diferentes tipos de amostras negativas, verifica-se que esse classificador apresentou melhores resultados nos cenários com amostras geradas por proximidade espacial.

Ainda, verificou-se que os algoritmos que utilizam o método de aprendizado *ensemble* (RF e LGBM) se destacaram, apresentando resultados superiores aos dos algoritmos tradicionais (RNA e AD).

Após a avaliação de performance dos modelos selecionados, observa-se que o LGBM apresenta uma escolha interessante como modelo final para predição de escorregamentos de encostas. Além de apresentar os melhores resultados, trata-se de um classificador com boa escalabilidade e, com exceção do AD, apresentou os menores tempos de execução. A RF pode ser um alternativa interessante no que tange ao desempenho preditivo, contudo obteve tempo de execução muito superior aos demais classificadores.

Tabela 4.2: Resultados da classificação para os cenários 1 e 2.

Cenário 1					
	F1	ACC	AUC	Recall	Precision
RNA	0,810±0,005	0,809±0,005	0,809±0,005	0,814±0,009	0,806±0,009
AD	0,889±0,003	0,885±0,003	0,885±0,003	0,914±0,008	0,865±0,005
RF	0,929±0,002	0,930±0,002	0,930±0,002	0,914±0,003	**0,944±0,004***
LGBM	**0,931±0,002***	**0,932±0,002***	**0,932±0,002***	**0,935±0,002***	0,930±0,002
Cenário 2					
RNA	0,811±0,005	0,811±0,005	0,811±0,005	0,815±0,006	0,809±0,008
AD	0,890±0,002	0,888±0,002	0,888±0,002	0,909±0,005	0,872±0,007
RF	0,929±0,003	0,931±0,003	0,931±0,003	0,913±0,004	**0,946±0,003***
LGBM	**0,934±0,002***	**0,934±0,002***	**0,934±0,002***	**0,937±0,003***	0,932±0,002

* Medida significativamente maior (Wilcoxon signed-rank, $\alpha < 0,05$)

Tabela 4.3: Resultados da classificação para os cenários 3 e 4.

Cenário 3					
	F1	ACC	AUC	Recall	Precision
RNA	0,815±0,004	0,815±0,004	0,814±0,004	0,819±0,006	0,812±0,006
AD	0,888± 0,003	0,885± 0,004	0,885± 0,004	0,909± 0,008	0,869± 0,007
RF	0,932± 0,002	0,933± 0,002	0,933± 0,002	0,916± 0,002	**0,948± 0,004***
LGBM	**0,935±0,001***	**0,935±0,001**	**0,935±0,001***	**0,937±0,002***	0,933±0, 002
Cenário 4					
RNA	0,814±0, 005	0,813±0, 005	0,813±0, 005	0,820±0, 008	0,810±0, 008
AD	0,887±0, 004	0,885±0, 004	0,885±0, 004	0,904±0, 005	0,871±0, 008
RF	0,931±0, 002	0,932±0, 002	0,932±0, 002	0,911±0, 002	**0,951±0,002***
LGBM	**0,937±0,002***	**0,937±0,002***	**0,937±0,002***	**0,940±0,002***	0,934±0, 004

* Medida significativamente maior (Wilcoxon signed-rank, $\alpha < 0,05$)

Tabela 4.4: Resultados da classificação para os cenários 5 e 6.

Cenário 5					
	F1	ACC	AUC	Recall	Precision
RNA	0,818±0, 004	0,818±0, 004	0,818±0, 004	0,818±0, 004	0,818±0, 004
AD	0,891±0, 003	0,889±0, 003	0,889±0, 003	0,913±0, 008	0,871±0, 005
RF	0,932±0, 003	0,933±0, 003	0,933±0, 003	0,919±0, 004	**0,946±0,003***
LGBM	**0,938±0,003***	**0,938±0,003***	**0,938±0,003***	**0,939±0,004***	0,938±0, 003
Cenário 6					
RNA	0,818±0, 002	0,817±0, 002	0,817±0, 002	0,822±0, 005	0,815±0, 005
AD	0,893±0, 004	0,891±0, 004	0,891±0, 004	0,911±0, 008	0,876±0, 007
RF	0,933±0, 002	0,935±0, 001	0,935±0, 001	0,915±0, 004	**0,952±0,004***
LGBM	**0,940±0,003***	**0,940±0,003***	**0,940±0,003***	**0,943±0,003***	0,937±0, 003

* Medida significativamente maior (Wilcoxon signed-rank, $\alpha < 0,05$)

Tabela 4.5: Resultados da classificação para os cenários 7 e 8.

Cenário 7					
	F1	ACC	AUC	Recall	Precision
RNA	0,863±0,002	0,866±0,002	0,866±0,002	0,841±0,005	0,886±0,005
AD	0,920±0,002	0,920±0,002	0,920±0,002	0,914±0,008	0,925±0,006
RF	0,938±0,002	0,939±0,002	0,939±0,002	0,926±0,003	**0,950±0,002***
LGBM	**0,942±0,002***	**0,942±0,002***	**0,942±0,002***	**0,937±0,003***	0,946±0,003
Cenário 8					
RNA	0,880±0,002	0,882±0,002	0,882±0,002	0,866±0,004	0,895±0,005
AD	0,919±0,003	0,920±0,003	0,920±0,003	0,907±0,006	0,932±0,006
RF	0,944±0,002	0,94±0,002	0,945±0,002	0,932±0,004	0,957±0,002
LGBM	**0,955±0,001***	**0,955±0,001***	**0,955±0,001***	**0,952±0,002***	**0,958±0,002**

* Medida significativamente maior (Wilcoxon signed-rank, $\alpha < 0,05$)

Tabela 4.6: Resultados da classificação para os cenários 9 e 10.

Cenário 9					
	F1	ACC	AUC	Recall	Precision
RNA	0,867±0,003	0,871±0,003	0,871±0,003	0,847±0,007	0,889±0,007
AD	0,925±0,002	0,925±0,003	0,925±0,003	0,921±0,008	0,929±0,009
RF	0,944±0,003	0,945±0,003	0,945±0,003	0,934±0,004	**0,955±0,003**
LGBM	**0,949±0,002***	**0,950±0,002***	**0,950±0,002***	**0,943±0,002***	0,956±0,002
Cenário 10					
RNA	0,886±0,004	0,888±0,004	0,888±0,004	0,873±0,006	0,900±0,003
AD	0,924±0,004	0,924±0,003	0,924±0,003	0,914±0,007	0,934±0,005
RF	0,947±0,002	0,948±0,002	0,948±0,002	0,937±0,003	0,959±0,003
LGBM	**0,956±0,002***	**0,957±0,002***	**0,957±0,002***	**0,953±0,003***	**0,960±0,002**

* Medida significativamente maior (Wilcoxon signed-rank, $\alpha < 0,05$)

Tabela 4.7: Resultados da classificação para os cenários 11 e 12.

Cenário 11					
	F1	ACC	AUC	Recall	Precision
RNA	0,871±0,003	0,874±0,003	0,874±0,003	0,852±0,004	0,892±0,007
AD	0,920±0,002	0,921±0,003	0,921±0,003	0,915±0,005	0,926±0,008
RF	0,945±0,002	0,946±0,002	0,946±0,002	0,934±0,004	0,957±0,003
LGBM	**0,950±0,001***	**0,950±0,001***	**0,950±0,001***	**0,944±0,002***	**0,956±0,001**
Cenário 12					
RNA	0,887±0,004	0,889±0,004	0,889±0,004	0,874±0,006	0,901±0,004
AD	0,926±0,002	0,926±0,002	0,926±0,002	0,919±0,008	0,933±0,006
RF	0,949±0,003	0,950±0,002	0,950±0,002	0,938±0,005	0,961±0,002
LGBM	**0,958±0,002***	**0,958±0,002***	**0,958±0,002***	**0,953±0,003***	**0,963±0,002**

* Medida significativamente maior (Wilcoxon signed-rank, $\alpha < 0,05$)

4.2 Resultados de Testes Adicionais

Para realização do teste adicional dos modelos, os classificadores treinados foram apresentados aos demais dados de teste separados no início, de maneira que os classificadores fossem utilizados em dados que nunca foram usados em momento anterior, inclusive não tendo sido usados para otimização de hiperparâmetros. Os resultados dos testes adicionais são aparentados nas Tabelas 4.8, 4.9, 4.10, 4.11, 4.12 e 4.13.

De maneira geral, pela comparação de valores obtidos com as médias anteriores, os modelos demonstraram estabilidade e foram capazes de repetir seus desempenhos nos dados adicionais, demonstrando boa capacidade de generalização. Novamente, o LGBM se destacou obtendo as melhores medidas de desempenho e sendo significativamente melhor na maioria dos cenários. Ainda pôde ser verificado que houve uma maior dispersão dos dados em relação à sua média, entretanto, esse aumento não compromete os bons resultados obtidos. Por fim, observou-se que o LGBM apresentou, nos cenários 3, 5 e 9, desempenho superior aos cenários em que não é considerado o dia posterior ao evento do escorregamento, diferentemente do que ocorreu nos primeiros resultados.

Tabela 4.8: Resultados de testes adicionais da classificação para os cenários 1 e 2.

Cenário 1					
	F1	ACC	AUC	Recall	Precision
RNA	0,810±0,011	0,810±0,011	0,810±0,011	0,805±0,023	0,813±0,011
AD	0,880±0,013	0,880±0,013	0,880±0,013	0,908±0,015	0,860±0,020
RF	0,932±0,005	0,932±0,005	0,932±0,005	0,917±0,010	**0,945±0,010***
LGBM	**0,933±0,005**	**0,933±0,005**	**0,933±0,005**	**0,939±0,010***	0,928±0,007
Cenário 2					
RNA	0,814±0,015	0,814±0,015	0,814±0,015	0,812±0,024	0,815±0,015
AD	0,884∓0,009	0,884∓0,009	0,884∓0,009	0,908∓0,013	0,866∓0,009
RF	0,927±0,005	0,927±0,005	0,927±0,005	0,911±0,010	**0,941±0,012***
LGBM	**0,936±0,005***	**0,936±0,005***	**0,936±0,005***	**0,943±0,007***	0,931±0,007

* Medida significativamente maior (Wilcoxon signed-rank, $\alpha < 0,05$)

Tabela 4.9: Resultados de testes adicionais da classificação para os cenários 3 e 4.

Cenário 3					
	F1	ACC	AUC	Recall	Precision
RNA	0,819±0,008	0,819±0,008	0,819±0,008	0,809±0,018	0,825±0,011
AD	0,875±0,007	0,875±0,007	0,875±0,007	0,908±0,013	0,852±0,014
RF	0,929±0,003	0,929±0,003	0,929±0,003	0,913±0,009	**0,942±0,008***
LGBM	**0,935±0,005**	**0,935±0,005**	**0,935±0,005**	**0,938±0,006***	0,932±0,006
Cenário 4					
RNA	0,821±0,010	0,821±0,010	0,821±0,010	0,824±0,015	0,820±0,020
AD	0,889±0,012	0,889±0,012	0,889±0,012	0,910±0,016	0,873±0,013
RF	0,936±0,007	0,936±0,007	0,936±0,007	0,913±0,011	**0,957±0,007***
LGBM	**0,939±0,005**	**0,939±0,005**	**0,939±0,005**	**0,939±0,010***	0,939±0,007

* Medida significativamente maior (Wilcoxon signed-rank, $\alpha < 0,05$)

Tabela 4.10: Resultados de testes adicionais da classificação para os cenários 5 e 6.

Cenário 5					
	F1	ACC	AUC	Recall	Precision
RNA	0,821±0,012	0,821±0,012	0,821±0,012	0,827±0,021	0,817±0,015
AD	0,885±0,008	0,886±0,008	0,886±0,008	0,916±0,011	0,863±0,011
RF	0,932±0,006	0,932±0,006	0,932±0,006	0,922±0,012	**0,941±0,007***
LGBM	**0,936±0,006***	**0,936±0,006***	**0,936±0,006***	**0,940±0,010***	0,934±0,006
Cenário 6					
RNA	0,818±0,010	0,818±0,010	0,818±0,010	0,835±0,020	0,808±0,020
AD	0,892±0,010	0,892±0,010	0,892±0,010	0,913±0,017	0,877±0,011
RF	0,934±0,007	0,934±0,007	0,934±0,007	0,915±0,015	**0,951±0,008***
LGBM	**0,938±0,009***	**0,938±0,009**	**0,938±0,009**	**0,940±0,016***	0,937±0,007

* Medida significativamente maior (Wilcoxon signed-rank, $\alpha < 0,05$)

Tabela 4.11: Resultados de testes adicionais da classificação para os cenários 7 e 8.

Cenário 7					
	F1	ACC	AUC	Recall	Precision
RNA	0,866±0,015	0,867±0,015	0,867±0,015	0,828±0,032	0,898±0,012
AD	0,915±0,004	0,915±0,004	0,915±0,004	0,910±0,010	0,921±0,011
RF	0,944±0,006	0,944±0,006	0,944±0,006	0,933±0,012	**0,954±0,013***
LGBM	**0,946±0,005***	**0,946±0,005***	**0,946±0,005***	**0,944±0,011***	0,947±0,010
Cenário 8					
RNA	0,889±0,006	0,889±0,006	0,889±0,006	0,869±0,014	0,905±0,012
AD	0,922±0,010	0,922±0,010	0,922±0,010	0,908±0,015	0,935±0,013
RF	0,946±0,004	0,946±0,004	0,946±0,004	0,936±0,008	0,954±0,004
LGBM	**0,955±0,006***	**0,955±0,006***	**0,955±0,006***	**0,953±0,008***	**0,957±0,006***

* Medida significativamente maior (Wilcoxon signed-rank, $\alpha < 0,05$)

Tabela 4.12: Resultados de testes adicionais da classificação para os cenários 9 e 10.

Cenário 9					
	F1	ACC	AUC	Recall	Precision
RNA	0,870±0,011	0,870±0,011	0,870±0,011	0,840±0,024	0,894±0,013
AD	0,926±0,006	0,926±0,006	0,926±0,006	0,920±0,011	0,932±0,012
RF	0,946±0,007	0,946±0,007	0,946±0,007	0,933±0,013	**0,957±0,006**
LGBM	**0,949±0,005***	**0,949±0,005***	**0,949±0,005***	**0,942±0,010***	0,955±0,006
Cenário 10					
RNA	0,889±0,007	0,889±0,007	0,889±0,007	0,870±0,011	0,905±0,013
AD	0,928±0,008	0,928±0,008	0,928±0,008	0,916±0,013	0,938±0,011
RF	0,950±0,008	0,950±0,008	0,950±0,008	0,941±0,010	0,958±0,009
LGBM	**0,955±0,005**	**0,955±0,005**	**0,955±0,005**	**0,950±0,010**	**0,959±0,005**

* Medida significativamente maior (Wilcoxon signed-rank, $\alpha < 0,05$)

Tabela 4.13: Resultados de testes adicionais da classificação para os cenários 11 e 12.

Cenário 11					
	F1	ACC	AUC	Recall	Precision
RNA	0,875±0,007	0,876±0,007	0,876±0,007	0,848±0,011	0,898±0,013
AD	0,925±0,007	0,925±0,007	0,925±0,007	0,919±0,012	0,930±0,010
RF	0,946±0,005	0,946±0,005	0,946±0,005	0,935±0,012	**0,956±0,006**
LGBM	**0,952±0,005***	**0,952±0,005***	**0,952±0,005***	**0,950±0,008***	0,954±0,008
Cenário 12					
RNA	0,892±0,009	0,892±0,009	0,892±0,009	0,879±0,013	0,903±0,011
AD	0,921±0,007	0,921±0,007	0,921±0,007	0,914±0,009	0,927±0,007
RF	0,949±0,006	0,949±0,006	0,949±0,006	0,941±0,010	0,956±0,006
LGBM	**0,958±0,006***	**0,958±0,006***	**0,958±0,006***	**0,957±0,008***	**0,959±0,008**

* Medida significativamente maior (Wilcoxon signed-rank, $\alpha < 0,05$)

4.3 Análise do número de dias de chuva mais efetivo na predição de escorregamentos

Com o objetivo de verificar o número de dias de chuva mais efetivo na predição de deslizamento de terra foram propostos mais quatro cenários de avaliação, tendo metade dos cenários com oito dias de chuva acumulada e o restante com quatro dias de chuva acumulada. Além disso, os cenários 13 e 15 foram desenvolvidos com amostras de um dia anterior ao evento do escorregamento e um dia de antecedência preditiva e os cenários 14 e 16 foram criados a partir de amostras negativas levando

em consideração o talude mais próximo e um dia de antecedência preditiva. Os resultados desses cenários são apresentados nas Tabelas 4.14 e 4.15. De modo geral, houve uma queda na performance dos classificadores a medida que foi diminuindo o número de dias de precipitação acumulada. Porém, os resultados para os cenários de 8 e 4 dias de chuva acumulada ainda são satisfatórios, alcançando valores médios de *F1-score* $= 0,877 \pm 0,003$ e AUC $= 0,875 \pm 0,003$ para 8 dias de chuva acumulada e valores médios de *F1-score* $= 0,788 \pm 0,002$ e AUC $= 0,788 \pm 0,005$ para 4 dias de precipitação acumulada.

Os algoritmos LGBM e RF demonstraram novamente os melhores resultados mesmo com a redução do número de atributos da base de dados. Contudo, diferentemente do que foi observado anteriormente, a RF apresentou significativamente os melhores resultados para esses cenários (Wilcoxon signed-rank, $\alpha <$ 0,05). Pode-se dizer que a redução do desempenho dos modelos está associada com a redução da dimensionalidade dos dados. Sendo assim, com mais informações relacionadas a precipitação pluviométrica os modelos são capazes de estimar com melhor eficácia a ocorrência de deslizamentos e não deslizamentos. Já os melhores resultados da RF em relação ao LGBM nesses cenários, pode estar associado, também, a capacidade dos classificadores em lidar com a dimensionalidade dos dados. Com isso, os resultados mostram que o LGBM é capaz alcançar melhores desempenhos, em relação a RF, em problemas com um maior número de variáveis.

Tanto o LGBM quanto o RF mantiveram a estabilidade entre a *recall* e a *precision*, além de apresentarem baixos valores de desvio padrão, denotando boa estabilidade desses classificadores.

Tabela 4.14: Resultados da classificação para os cenários 13 e 14.

Cenário 13					
	F1	ACC	AUC	Recall	Precision
RNA	0,770±0,008	0,763±0,013	0,763±0,013	0,792±0,037	0,752±0,029
AD	0,843±0,002	0,829±0,004	0,829±0,004	**0,913±0,012**	0,783±0,009
RF	**0,877±0,003**	**0,875±0,003***	**0,875±0,003***	0,889±0,004	**0,865±0,004***
LGBM	0,875±0,003	0,871±0,003	0,871±0,003	0,899±0,003*	0,852±0,004
Cenário 14					
RNA	0,781±0,004	0,791±0,005	0,791±0,005	0,746±0,005	0,821±0,011
AD	0,843±0,002	0,837±0,002	0,837±0,002	0,877±0,046	0,817±0,042
RF	**0,877±0,003***	**0,880±0,003***	**0,880±0,003***	0,859±0,005	**0,897±0,006***
LGBM	0,871±0,003	0,870±0,003	0,870±0,003	**0,875±0,005***	0,867±0,003

* Medida significativamente maior (Wilcoxon signed-rank, $\alpha < 0,05$)

Tabela 4.15: Resultados da classificação para os cenários 15 e 16.

Cenário 15					
	F1	ACC	AUC	Recall	Precision
RNA	0,734±0,019	0,689±0,013	0,689±0,013	0,863±0,054	0,641±0,009
AD	0,778±0,003	0,733±0,004	0,733±0,004	**0,935±0,003**	0,666 ±0,004
RF	**0,788±0,002***	0,750±0,003	0,750±0,003	0,929±0,001*	0,684±0,016
LGBM	0,785±0,002	**0,755±0,003***	**0,755±0,003***	0,895±0,0,007	**0,700±0,005***
Cenário 16					
RNA	0,700±0,010	0,667±0,031	0,667±0,031	0,786±0,103	0,651±0,069
AD	0,758±0,003	0,699±0,005	0,699±0,005	**0,945±0,009**	0,634±0,006
RF	**0,786±0,003***	**0,788±0,005***	**0,788±0,005***	0,780±0,007	**0,794±0,010***
LGBM	0,765±0,002	0,755±0,005*	0,755±0,005*	0,798±0,013*	0,736±0,012

* Medida significativamente maior (Wilcoxon signed-rank, $\alpha < 0,05$)

4.4 Importância dos atributos

Na Figura 4.1 e 4.2, são apresentados os ranqueamentos dos atributos usados pelo LGBM. Esse ranqueamento é baseado no índice composto com GINI e na quantidade de árvores onde o atributo é utilizado. Este valor é calculado automaticamente para cada atributo após a etapa de treinamento do LGBM e pode-se observar que segundo esse índice, os melhores preditores são os atributos pluviométricos diários associados a cada deslizamento de terra. Isso ocorre tanto para o ranqueamento dos atributos na base de dados com amostras negativas por proximidade temporal (Figura 4.1), quanto para o ranqueamento dos atributos na base de dados com amostras negativas por proximidade espacial (Figura 4.2).

Ao analisar os dois ranqueamentos, nota-se um diferença na ordem dos atributos em função do tipo de amostras negativas. Apesar das diferenças, as precipitações diárias alternam-se entre os atributos mais importantes para os modelos. Além disso, é possível verificar que os atributos relacionados às características geológicas apresentam menor importância para predição de deslizamentos de terra nas duas bases de dados analisadas. Desta forma, a partir da análise da importância dos atributos pode-se buscar identificar um conjunto menor de atributos mais relevantes, que possam simplificar o modelo, mantendo ou melhorando o desempenho da predição.

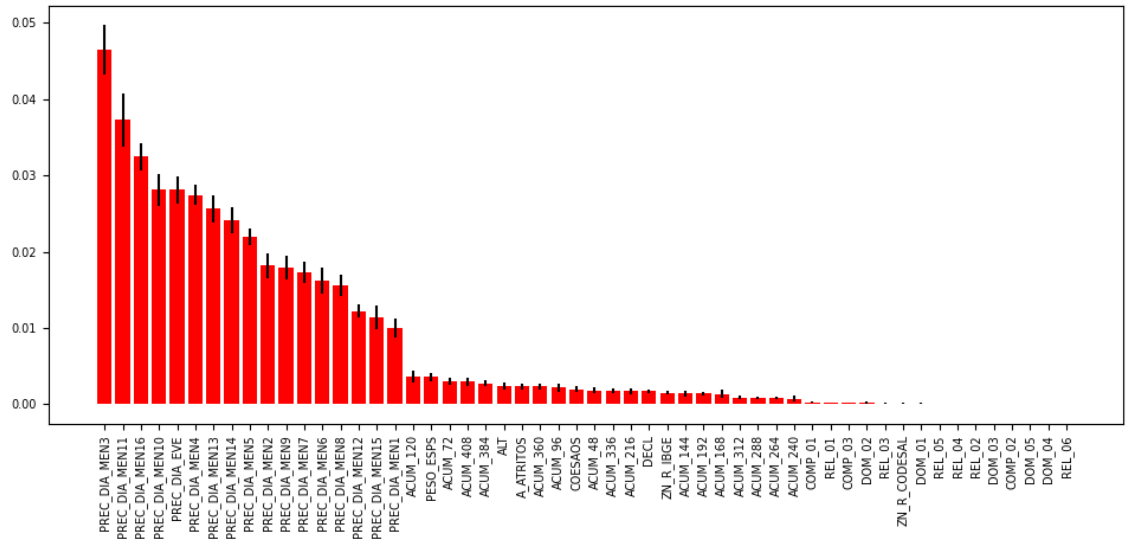

Figura 4.1: Ranqueamento de atributos por importância pelo LGBM com amostras negativas com proximidade temporal.

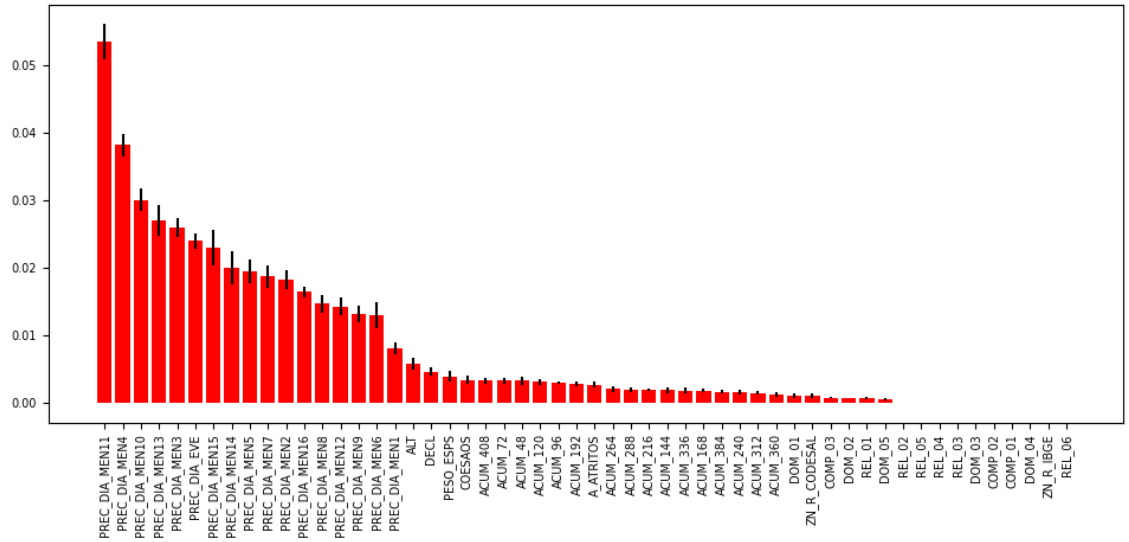

Figura 4.2: Ranqueamento de atributos por importância pelo LGBM com amostras negativas com proximidade espacial.

Devido às diferenças identificadas com o ranqueamento do LGBM, foram realizados testes com conjuntos de atributos menores (30, 25, 20 e 15), selecionados pelo grau de importância, e comparados nas tabelas 4.16 e 4.17, buscando identificar quais atributos produzem melhores resultados para predição de deslizamentos de terra.

Tabela 4.16: Resultados do LGBM com amostras negativas com proximidade temporal 15, 20, 25 e 30 atributos, selecionados pelo ranking de atributos.

	F1	ACC	AUC	Recall	Precision
15 atributos	0,931±0, 002	0,931±0, 002	0,931±0, 002	0,935±0, 003	0,928±0, 003
20 atributos	0,934±0, 002	0,934±0, 002	0,934±0, 002	0,936±0, 003	0,931±0, 003
25 atributos	0,933±0, 002	0,932±0, 002	0,932±0, 002	0,935±0, 002	0,931±0, 002
30 atributos	0,932±0, 002	0,932±0, 002	0,932±0, 002	0,936±0, 002	0,929±0, 002

Tabela 4.17: Resultados do LGBM com amostras negativas com proximidade espacial 15, 20, 25 e 30 atributos, selecionados pelo ranking de atributos.

	F1	ACC	AUC	Recall	Precision
15 atributos	0,940±0, 002	0,940±0, 002	0,940±0, 002	0,937±0, 002	0,943±0, 004
20 atributos	0,942±0, 002	0,943±0, 002	0,943±0, 002	0,938±0, 003	0,947 ±0, 003
25 atributos	0,944±0, 002	0,944±0, 002	0,944±0, 002	0,940±0, 003	0,948±0, 003
30 atributos	0,944±0, 003	0,944±0, 003	0,944±0, 003	0,939±0, 003	0,949±0, 004

Conforme pode ser observado na Tabela 4.16, a base de dados composta por amostras negativas com proximidade temporal apresentou os melhores resultados com os 20 atributos mais bem ranqueados. Já de acordo com a Tabela 4.17, para base de dados com amostras negativas com proximidade espacial, a diminuição do número de atributos mais importantes teve como consequência a redução dos valores de todas as medidas de avaliação do LGBM. Os modelos ainda mantiveram equilíbrio entre a *recall* e a *precision* e baixo desvio padrão. Os resultados indicam que podem ser realizadas melhorias na seleção de atributos que possam reduzir o número de atributos coletados sem causar diferenças relevantes, mantendo a predição de deslizamentos de terra confiável.

4.5 Discussão

É sabido que o desempenho de um modelo de aprendizado de máquina depende do conhecimento pertinente ao problema que se pretende analisar e da disponibilidade de dados com variáveis informativas e com capacidade discriminatória em relação à variável de resposta. Desse modo, objetivando construir uma base de dados com características essenciais para o desenvolvimento de modelos de predição de escorregamento de encostas, foram integrados diferentes conjuntos de dados com preditores geológicos, geomorfométricos e precipitação pluviométrica, conforme indicado em literatura relacionada aos movimentos de massa e a predição de deslizamento de

terra. De posse da base de dados e tendo realizado as etapas de pré-processamento, foram aplicados algoritmos de aprendizado de máquina para aprender e identificar padrões nesses dados, a fim de realizar a predição.

De modo geral, todos os classificadores avaliados alcançaram performances expressivas e promissoras, sendo os melhores resultados observados para o LGBM, que obteve valores médios de *F1-score* superiores a $(0,929 \pm 0,002)$ e de AUC superiores a $(0,930 \pm 0,002)$, para os cenários com 16 dias de chuva acumulada. As pequenas dispersões nos resultados e o desempenho repetido sobre o conjunto de testes adicionais demonstram que os modelos foram capazes de aprender e otimizar a predição mesmo com diferentes conjuntos de dados.

As diferenças de desempenho dos classificadores podem estar relacionadas às características dos dados, à capacidade de aprendizado dos algoritmos, às técnicas e métodos utilizados para construir e avaliar os modelos, como o aperfeiçoamento dos hiperparâmetros dos algoritmos. Já as diferenças de performance dos modelos preditivos em cada cenário de avaliação são atribuídas às diferentes técnicas de geração de amostras negativas, à quantidade de dias de chuva acumulada considerada e aos diferentes espaços temporais de antecedência preditiva.

O melhor desempenho do LGBM nos cenários com 16 dias de chuva acumulada, mostra uma melhor robustez desse algoritmo em conjunto de dados com grande número de variáveis. Já o melhor desempenho da RF em cenários com 8 e 4 dias de chuva acumulada, mostra a necessidade de se realizar uma seleção de atributos na tentativa de obter um melhor o desempenho desse classificador em cenários com maior número de variáveis.

Para a implementação de modelos preditivos de escorregamentos de encostas, casos de não escorregamentos precisaram ser adicionados ao conjunto de amostras, tendo assim, duas classes: positiva (escorregamento) e negativa (não escorregamento). Entretanto, é importante ressaltar que a depender do método aplicado, as amostras negativas podem ser facilmente classificáveis, inflando as métricas de avaliação. Um exemplo disso, é gerar amostras negativas com grande intervalo de tempo em relação à ocorrência do evento, ou gerar amostras a partir de locais em que a ocorrência de um escorregamento seja improvável.

Com base nessas considerações, foram propostas duas metodologias: uma baseada nos dados da própria encosta de referência e a outra baseada na encosta mais próxima. Dentre as duas técnicas de geração de amostras negativas avaliadas, observou-se um melhor desempenho dos classificadores sobre a base com amostras negativas geradas com proximidade espacial. Dessa forma, foi entendido que os classificadores conseguiram aprender e identificar padrões que relacionavam os fatores condicionantes.

A classificação em cima da base de dados com amostras geradas por proximidade temporal, também apresentou resultados expressivos e expõe a capacidade dos modelos em identificar padrões que correlacionam o desencadeamento da ocorrência de

um escorregamento com os consecutivos dias de variadas precipitações, mesmo que o limite entre a estabilidade das encostas (fatores de controle) e o quanto choveu em determinado tempo seja pequeno. Desta forma, podemos inferir que o método de geração de amostras negativas pode impactar a capacidade preditiva do modelo.

Com o objetivo de verificar quantos dias de chuva acumulada são mais efetivos na predição de escorregamentos, foram experimentados os valores de 4, 8 e 16 dias. Verificou-se que a capacidade preditiva dos algoritmos aumentou com o acréscimo do número de dias de chuva acumulada, o que sugere uma relação entre o número de ocorrências de deslizamentos e diferentes períodos de chuva acumulada. Para 16 dias, os algoritmos apresentaram os melhores desempenhos, sendo mais eficazes na predição. Os cenários de 4 e 8 dias demonstraram que as informações de precipitação desse período são bastantes relevantes, visto que já permitiram obter modelos com performances razoáveis. Com isso, nota-se que os modelos preditivos foram capazes de identificar os eventos de deslizamentos de terra desencadeados por chuvas intensas em um curto período, bem como deslizamentos de terra ativados por chuvas de intensidade pequena a moderada, mas constante por um longo período.

Também foram experimentados diferentes espaços temporais de antecedência de previsão, a fim de verificar a implicação destes na capacidade preditiva dos modelos. Foram considerados três espaços temporais: 1, 2 e 3 dias de antecedência. Ao analisar os cenários sob a óptica da antecedência de previsão verificou-se que, diferentemente do que se esperava, houve um pequeno acréscimo na capacidade preditiva dos algoritmos com o aumento da antecedência da previsão da precipitação. O motivo desse acréscimo pode estar relacionado com a utilização dos dados reais de precipitação ao invés da previsão de precipitação propriamente dita, normalmente sujeita à imprecisões. Com isso, o acréscimo na performance dos classificadores decorre do aumento do número de dias de chuva acumulada, ou seja, devido ao aumento da janela temporal de dados de precipitação. Desta forma, para uma melhor compreensão da influência da predição de chuvas na antecedência preditiva dos modelos estudados, novos estudos devem ser realizados, a fim de verificar o impacto das incertezas da previsão de precipitação na capacidade preditiva dos modelos propostos.

Os cenários 2, 4, 6, 8, 10 e 12, consideram dados de precipitação do dia posterior ao escorregamento devido à imprecisão do registro do momento do escorregamento (data e hora). Esses cenários apresentam, na maioria das situações, um acréscimo de desempenho em relação aos cenários que não consideram o dia posterior ao escorregamento. Entende-se que esse acréscimo ocorreu devido ao aumento do número de dias de chuva e de seus respectivos acumulados e a possíveis casos de escorregamentos que tenham sido registrados até um dia depois da sua ocorrência.

Considerando o ranqueamento dos atributos, os expressivos resultados com um número significativamente menor de atributos permite concluir que é viável realizar a predição de deslizamentos de terra apenas com as informações de inventários de deslizamentos de terra e precipitações pluviométricas, o que simplifica a implementação de modelos de predição de deslizamentos de terra baseada em aprendizado de

máquina.

De modo comparativo, em relação aos trabalhos relacionados (Souza e Ebecken, 2012; Farahmand e Aghakouchak, 2013; Tehrani et al., 2019), os modelos propostos apresentaram melhores resultados numéricos absolutos. Contudo, como os modelos foram construídos sob dados com características distintas e foram empregados diferentes métodos de criação de amostras negativas, este comparativo torna-se ilustrativo, sem significativo rigor científico.

Avaliando os resultados obtidos em geral, o LGBM demonstrou melhor adaptação aos dados apresentados, como visto através dos melhores valores de F1-score e AUC baixa dispersão em relação a média e valores de *precision* e *recall* equilibrados, isso devido a não-linearidade dos parâmetros, homogeneidade da variância e independência entre os atributos preditores. Assim, considera-se que, sendo aplicada a base com amostras negativas com proximidade espacial e 16 dias de chuva acumulada, o LGBM, nestas condições, obteve melhor desempenho se comparado com os demais modelos ou tipo de amostras negativas e dias de chuva acumulada.

Os resultados aqui apresentados são animadores, com isso, espera-se que, a partir da metodologia proposta, seja possível contribuir com o desenvolvimento da área de pesquisa e o desenvolvimento de aplicações práticas. Mostramos o grande potencial da aplicação de mineração de dados e aprendizado de máquina na predição de deslizamentos de terra e, apesar das limitações, em última instância, a implementação dos modelos propostos já pode servir como ferramenta para auxiliar o processo de tomadas de decisões possivelmente reduzindo os danos causados pelos deslizamentos de terra.

4.5.1 Limitações e Direcionamentos

Além do que foi discutido até o momento em relação aos resultados, é importante abordar algumas dificuldades e limitações acerca dos dados utilizados, bem como sugerir melhorias no processo de coleta de dados, pois a qualidade da coleta reflete na qualidade dos dados, podendo trazer benefícios para o processo de aprendizado, e consequentemente melhorar a capacidade preditiva dos algoritmos.

De forma a contornar algumas limitações nos dados, algumas simplificações foram realizadas. Nos dados geotécnicos ou das propriedades do solo foi adotado tanto para as amostras de escorregamentos e de não escorregamentos das propriedades do solo (coesão, ângulo de atrito e peso específico) na condição próxima à saturação. Essa simplificação se deu sobretudo devido à indisponibilidade de dados, não sendo possível estabelecer uma relação ideal entre a pluviosidade e as alterações nos valores das propriedades do solo, pois sabe-se que um aumento do teor de umidade dissolve os agentes cimentantes, reduzindo a coesão e o ângulo de atrito, podendo reduzir as tensões de sucção causando bruscas reduções de volume e colapsando o solo. Para estabelecer o real valor dessas propriedades no momento do escorregamento e do não escorregamento seria necessário dispor da curva de retenção do solo (relação entre a

sucção e a variação do teor de umidade do solo) de cada encosta e da relação entre a precipitação e o teor umidade do solo (coeficiente de permeabilidade). A disponibilidade dessas informações também permitiria considerar e identificar a ocorrência de escorregamento em condições de solos não saturados. Não sendo possível estabelecer essas relações, adotou-se a mesma situação para deslizamento e não deslizamento.

Além das simplificações, algumas limitações foram observadas nos conjuntos de dados utilizados na construção do *dataset* e na aplicação de algoritmos de aprendizado de máquina. Os dados das propriedades de solo foram obtidos por meio da interpolação de informações extraídas de ensaios laboratoriais, onde foram utilizadas 385 amostras de blocos indeformados de solos, coletados em diversos pontos do município de Salvador. A metodologia adotada já apresenta aproximações em sua essência. Dado que ensaiar uma amostra de solo para cada encosta é quase impraticável, para que se tenha dados com maior precisão, sugere-se a adição de amostras de solos distribuídas estrategicamente ao longo da região estudada, incluindo, principalmente, as áreas onde são frequentes os eventos de deslizamentos de terra.

No que tange aos dados geomorfométricos, o MDE utilizado nesta pesquisa possui resolução espacial de 30m. O desenvolvimento de modelos de alta resolução espacial irá contribuir para uma maior precisão na atribuição da declividade e da altitude para cada encosta.

Os dados de inventários de deslizamentos também apresentam algumas limitações. Como dito anteriormente, a data e hora do processo não consistem no exato momento que ocorreu o escorregamento. Com isso, de modo a atenuar essa imprecisão, sugere-se que ao realizar o registro da ocorrência seja registrada também o horário, mesmo que aproximado, do momento do colapso. De posse desse registro seria possível a utilização dos dados de precipitação pluviométricas com menor granularidade, por exemplo, intervalos de precipitação a cada hora ou a cada 30 minutos. Em consequência disso, as amostras de não deslizamentos também poderiam ser criadas com menor espaço temporal anterior ao momento do evento crítico.

Além do problema do momento do escorregamento, verificou-se que não foi adotada uma padronização completa na coleta de dados durante a realização da vistoria técnica no local onde houve o escorregamento. A falta de padronização impede que algumas informações relevantes sejam associadas aos registros de escorregamento. Isso ocorre pois as informações são apresentadas em formato de texto livre, onde cada profissional pode atribuir uma nomenclatura para qualificar ou descrever determinada característica da encosta ou ocorrência. A simples atribuição da causa de escorregamento em forma de texto livre dificulta a separação dos eventos de escorregamentos induzidos por chuva e dos deslizamentos causados, por exemplo, pelo rompimento de uma tubulação. Além disso, informações contidas nas descrições como, porte do deslizamento, tipo de vegetação entre outras, tornam-se de difícil uso, visto que não existe uma padronização que categorize essas informações.

Capítulo 5

Conclusões

A implementação de um modelo de escorregamentos de encostas é uma tarefa complexa devido ao grande número de variáveis envolvidas e a enorme variedade (espacial e temporal) dos parâmetros. Sendo assim, esta pesquisa apresentou como objetivo propor e avaliar a aplicação de métodos de mineração de dados e aprendizado de máquina, no pré-processamento e na classificação utilizando Redes Neurais Artificiais (RNA) do tipo *Multi-Layer Perceptron*, *Random Forest* (RF), Árvore de Decisão (AD) e *Light Gradient Boosting Machine* (LGBM), para predição de escorregamentos de encostas e, para isso, foi necessária a construção de uma base de dados a partir da integração da informações de múltiplas fontes.

Os resultados obtidos da AUC, ACC, *F1-score*, *precison* e *recall* de cada algoritmo de classificação foram analisados com o intuito de identificar qual destes obteria melhor performance. Também foram avaliados diversos cenários de observação com o intuito de verificar a influência do método de geração de amostras negativas, a interferência do número de dias de chuva acumulada e o impacto da antecedência de previsão na capacidade preditiva dos classificadores.

A partir dos resultados foi possível concluir, em linhas gerais, que os modelos de classificação são ferramentas capazes de encontrar padrões na base de dados proposta e que podem, com alta performance preditiva, realizar a previsão de deslizamentos de terra. Além disso, reconhecer algumas limitações e imprecisões nos dados permite também estimar que com dados com mais qualidade a capacidade preditiva dos algoritmos melhore.

Devido às características de não-linearidade do problema e, consequentemente, da base de dados utilizada, o RF e o LGBM obtiveram melhores desempenhos em relação ao AD e à RNA. Dentre todos os classificadores, o LGBM apresentou as melhores medidas de avaliação em quase todos dos cenários de análise. Além disso, obteve em todos os cenários as melhores medidas de *recall*, o que denota uma maior capacidade deste algoritmo em identificar corretamente a maioria dos casos de interesse.

Os resultados desse estudo de predição de escorregamentos confirmaram a impor-

tância do método de geração de amostras negativas, visto que houve diferenças preditivas entre as duas amostras adotadas. Com isso, os classificadores obtiveram melhores resultados quando aplicados nas bases de dados em que as amostras negativas foram criadas a partir da encosta mais próxima que deslizou em momento posterior.

Este estudo confirmou também a importância da chuva acumulada e o número de dias efetivos no processo de desencadeamento de deslizamentos de terra, ficando estabelecido como dezesseis o número de dias de chuva acumulada que apresentou melhores resultados preditivos. Entretanto, os resultados dos modelos indicam que as precipitações entre quatro e oito dias são as informações mais relevantes no processo de desencadeamento de escorregamentos e que o acréscimo de dias confere um refinamento e um ganho de performance dos classificadores, possivelmente devido à influência das chuvas de longo período no processo de desencadeamento do escorregamentos de encostas.

No que tange à antecedência preditiva, não foi possível determinar qual o impacto do tempo de antecedência de previsão de precipitação na performance dos algoritmos, visto que foram utilizados dados reais de precipitação, devido à indisponibilidade de históricos de previsão de precipitação.

A análise da utilização de métodos de classificadores *ensembles* confirmaram seu potencial em conferir melhores resultados preditivos se comparados aos classificadores tradicionais, visto que o RF e o LGBM obtiveram as melhores performances em todos os cenários analisados.

Desta forma, a qualidade dos resultados obtidos para a predição de escorregamento de encostas induzidos por chuvas indica que a inserção desses modelos permitirá que a tomada de decisão, pelos órgãos competentes, seja mais embasada, conferindo melhorias no processo de monitoramento e redução dos dados causados pelos escorregamentos de encostas, de forma que podem ser utilizados nos sistemas de alarme existentes.

5.1 Trabalhos Futuros

Apesar dos resultados promissores, melhorias na metodologia podem ser consideradas. Cenários com um número substancialmente maior de amostras negativas devem ser analisados, visto que o estudo de escorregamentos de encosta está relacionado à identificação de eventos raros, ou seja, existe um número pequeno de ocorrência de deslizamentos de terra se comparado ao grande número de não ocorrências em um determinado período, mesmo durante eventos de chuva. Além disso, pode-se experimentar novos métodos de geração de amostras negativas, bem como a combinação de dois ou mais métodos.

Outra análise importante, que não foi realizada em virtude da inexistência de históricos de previsão de precipitação para a região estudada, é a verificação da sensi-

bilidade dos modelos preditivos propostos no que se refere à imprecisão da previsão de precipitação.

Com a instalação das estações geotécnicas, pode-se utilizar os dados por elas coletados para entender melhor as relações entre as variáveis geotécnicas (coesão, ângulo de atrito e peso específico) com a precipitação e com os escorregamentos de encostas.

Por fim, podem ser exploradas novas técnicas, como a utilização de aprendizado profundo à medida que se tem uma maior disponibilidade de dados, séries temporais para os dados de precipitação pluviométricas, e a construção de protótipos preditivos com o objetivo de validar, com o suporte de especialistas, os modelos propostos.

Referências

Achour, Y. e Pourghasemi, H. R. (2020). How do machine learning techniques help in increasing accuracy of landslide susceptibility maps? *Geoscience Frontiers*, 11(3):871–883.

Almeida, M., Nakazawa, A., e Tatizana, C. (1991). Análise de correlação entre chuvas e escorregamentos no município de petrópolis, rj. *Anais do 7o Congr. Bras. Geol. Engenharia, ABGE*, páginas 129–137.

Amado, A. (2015). Sobe para 15 número de mortes causadas por temporal em salvador. Disponível em: https://agenciabrasil.ebc.com.br/geral/noticia/2015-04/sobe-para-15-mumero-de-mortes-do-temporal-em-salvador. Acesso em: 02 de Fevereiro de 2022.

Aristizábal, E. e Sánchez, O. (2020). Spatial and temporal patterns and the socioeconomic impacts of landslides in the tropical and mountainous Colombian Andes. *Disasters*, 44(3):596–618.

Augusto Filho, O. (1992). Caracterização geológico-geotécnica voltada à estabilização de encostas: uma proposta metodológica MST. *: Conferência Brasileira sobre Estabilidade de Encostas*, 1:721–733.

Augusto Filho, O. e Virgili, J. C. (1998). Estabilidade de taludes. *Geologia de Engenharia. ABGE, São Paulo*, páginas 243–269.

Baum, R. L., Savage, W. Z., e Godt, J. W. (2008). TRIGRS — A Fortran Program for Transient Rainfall Infiltration and Grid-Based Regional Slope-Stability Analysis, Version 2.0. *U.S. Geological Survey Open-File Report*, (2008-1159):75.

Bergstra, J. e Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer.

Brand, E. W., Premchitt, J., e Phillipson, H. (1984). Relationship between rainfall and landslides in hong kong. In *Proceedings of the 4th International Symposium on Landslides*, volume 1, páginas 276–84. Canadian Geotechnical Society Toronto.

Brasil (2008). *Instituto Nacional de Pesquisas Espaciais (INPE). Topodata: banco de dados geomorfométricos do Brasil. Variáveis geomorfométricas locais.*

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

Breiman, L. (2004). Consistency for a simple model of random forests.

Breiman, L., Friedman, J. H., Olshen, R. A., e Stone, C. J. (1984). Classification and regression trees. belmont, ca: Wadsworth. *International Group*, 432:151–166.

Brink, H., Richards, J., e Fetherolf, M. (2016). *Real-World Machine Learning*. Manning Publications.

Brito, M. d. (2014). *Geoprocessamento aplicado ao mapeamento da suscetibilidade a escorregamentos no município de Porto Alegre, RS (Geoprocessing applied for landslide susceptibility mapping in the Porto Alegre municipality, RS, Brazil)*. Tese de Doutorado.

Campbell, R. H. (1975). *Soil slips, debris flows, and rainstorms in the Santa Monica Mountains and vicinity, southern California*, volume 851. US Government Printing Office.

Can, A., Dagdelenler, G., Ercanoglu, M., e Sonmez, H. (2017). Landslide susceptibility mapping at Ovacık-Karabük (Turkey) using different artificial neural network models: comparison of training algorithms. *Bulletin of Engineering Geology and the Environment*, 78(1):89–102.

Carson, M., Carson, M., Carson, M., e Kirkby, M. (1972). *Hillslope Form and Process*. Número v. 10 in Cambridge Geographical Studies. Cambridge University Press.

Castro, J. M. G. (2006a). Pluviosidade e movimentos de massa nas encostas de Ouro Preto. PÓS – GRAD:154.

Castro, J. M. G. (2006b). Pluviosidade e movimentos de massa nas encostas de ouro preto.

Chakraborty, A. e Goswami, D. (2017). Prediction of slope stability using multiple linear regression (MLR) and artificial neural network (ANN). *Arabian Journal of Geosciences*, 10(17):1–11.

Christofoletti, A. (1980). *Geomorfologia*, volume 2ed. Edgar Bucher, São Paulo.

Corani, G., Benavoli, A., Demšar, J., Mangili, F., e Zaffalon, M. (2017). Statistical comparison of classifiers through Bayesian hierarchical modelling. *Machine Learning*, 106(11):1817–1837.

Cruden, D. e Varnes, D. (1996). Landslides: Investigation and Mitigation. Chapter 3 - Landslide Types and Processes. *Transportation Research Board Special Report*, (247).

de Sousa Pinto, C. (2016). *Curso básico de Mecânica dos Solos*. Oficina de Textos.

D'Orsi, R. (2011). Correlação entre pluviometria e escorregamentos no trecho da Serra dos Órgãos da rodovia federal BR-116 RJ (Rio-Teresópolis). *Coppe - Ufrj*, página 303 pages.

Dou, J., Yunus, A. P., Tien Bui, D., Sahana, M., Chen, C.-W., Zhu, Z., Wang, W., e Pham, B. T. (2019). Evaluating GIS-Based Multiple Statistical Models and Data Mining for Earthquake and Rainfall-Induced Landslide Susceptibility Using the LiDAR DEM. *Remote Sensing*, 11(6):638.

Elbachá, A. T., Campos, L. E. P., e Bahia, R. F. C. (1992). Tentativa de correlação entre preciptação e deslizamento na cidade de salvador. *I Conferência Brasileira sobre Estabilidade de Encostas*, III.

Endo, T. (1969). Probable distribution of the amount of rainfall causing landslides, annual report 1968, hokkaido branch. *For. Exp. Stn., Sapporo, Japan*, páginas 122–136.

Eyles, R. (1979). Slip-triggering rainfalls in wellington city, new zealand.

Faceli, K., Gama, J., Lorena, A., e De Carvalho, A. (2011). *Inteligência artificial: uma abordagem de aprendizado de máquina*. Grupo Gen - LTC.

Farah, F. (2003). *Habitação e encostas*, volume volume. Instituto de Pesquisas Tecnológicas (IPT), São Paulo.

Farahmand, A. e Aghakouchak, A. (2013). A satellite-based global landslide model. *Natural Hazards and Earth System Science*, 13(5):1259–1267.

Fayyad, U., Piatetsky-Shapiro, G., e Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37.

Fernandes, N. e Amaral, C. (1996). Movimentos de massa: uma abordagem geológico-geomorfológica.(1996). *Geomorfologia e Meio Ambiente. Bertrand, Rio de Janeiro*, páginas 123–1994.

Fernandes, N. F. e Amaral, C. d. (2000). Movimentos de massa: uma abordagem geológico-geomorfológica. *Geomorfologia e Meio Ambiente. Bertrand, Rio de Janeiro*, páginas 123–194.

Fernandes, N. F., Guimarães, R. F., Gomes, R. A. T., Vieira, B. C., Montgomery, D. R., e Greenberg, H. (2001). Condicionantes Geomorfológicos dos Deslizamentos nas Encostas: Avaliação de Metodologias e Aplicação de Modelo de Previsão de Áreas Susceptíveis. *Revista Brasileira de Geomorfologia*, 2(1).

Ferrari, D. e De Castro, L. (2017). *Introdução a mineração de dados*. Saraiva Educação S.A.

Finlay, P., Fell, R., e Maguire, P. (1997). The relationship between the probability of landslide occurrence and rainfall. *Canadian Geotechnical Journal*, 34(6):811–824.

Freire, E. S. M. (1965). Movimentos coletivos de solos e rochas e sua moderna sistemática MST. *Construção*, páginas 10–8.

GEO-RIO (2000). Manual Técnico de Encostas. *Geo-Rio / Pcrj*, I:520.

Goldschmidt, R. e Passos, E. (2005). *Data mining: um guia Prático*. Elsevier Editora.

Govi, M. (1977). Photo-interpretation and mapping of the landslides triggered by the friuli earthquake (1976). *Bulletin of the International Association of Engineering Geology-Bulletin de l'Association Internationale de Géologie de l'Ingénieur*, 15(1):67–72.

Gramani, M. e Kanji, M. (2001). Inventário e análise das corridas de detritos no brasil. In *CONFERÊNCIA BRASILEIRA DE ESTABILIDADE DE ENCOSTAS*, volume 3, páginas 53–60.

Guidicini, G. e Iwasa, O. (1977). Tentative correlation between rainfall and landslides in a humid tropical environment. *Bulletin of the International Association of Engineering Geology-Bulletin de l'Association Internationale de Géologie de l'Ingénieur*, 16(1):13–20.

Guidicini, G. e Nieble, C. M. (1984). *Estabilidade de Taludes Naturais e de Escavação*, volume 2ed. Edgard Blücher, São Paulo.

Guimarães, R. F., Fernandes, N. F., Gomes, R. A. T., e Júnior, O. A. d. C. (2003). Fundamentação teórica do modelo matemático para previsão de escorregamentos rasos shallow stability. *Espaço Geografia*, 3(3).

Han, J., Kamber, M., e Pei, J. (2012). Data mining concepts and techniques, third edition.

Han, J., Pei, J., e Kamber, M. (2011). *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science.

Hanifinia, A., Nazarnejad, H., Najafi, S., Kornejady, A., e ... (2021). Landslide susceptibility assessment and mapping using statistical and data mining models in Iran.

Hastie, T., Tibshirani, R., e Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edição.

Haykin, S. S. (2009). *Neural networks and learning machines*. Pearson Education, third edição.

Henrique, A. (2022). Em 1 mês, sp tem mais mortes por deslizamentos que em todo o ano de 2021. Disponível em: https://www1.folha.uol.com.br/cotidiano/2022/02/em-1-mes-sp-tem-mais-morte-por-deslizamento-que-em-todo-o-2021.shtml. Acesso em: 02 de Fevereiro de 2022.

Highland, L. M. e Bobrowsky, P. (2008). O manual de deslizamento - um guia para a compreensão de deslizamentos. *US Geological Survey Circular*, 1325:156.

IBGE (2019). *Suscetibilidade a deslizamentos do Brasil: primeira aproximação*. Instituto Brasileiro de Geografia e Estatística - IBGE, Rio de Janeiro.

Ide, F. S. (2005). *Escorregamento, meteorologia e precipitação: uma proposta de método de investigação para a prevenção e monitoramento de riscos, aplicado em Campinas/SP*. Instituto de Pesquisas Tecnológicas do Estado de São Paulo.

IPT (1991). *Ocupação de Encostas*. IPT, São Paulo.

James, G., Witten, D., Hastie, T., e Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R*. Spinger.

Jesus, A. C. (2008). Retroanálise de Escorregamentos em Solos Residuais não Saturados. página 284.

Kay, J. e Chen, T. (1995). Rainfall-landslide relationship for hong kong. *Proceedings of the Institution of Civil Engineers-Geotechnical Engineering*, 113(2):117–118.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., e Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 2017-December(Nips):3147–3155.

Kelleher, J. D., Namee, B. M., e D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics*. Número 1.

Knust, K. (2021). Tragédia na Serra do RJ completa dez anos; R$ 500 milhões ainda devem ser investidos em obras. Disponível em: https://g1.globo.com/rj/regiao-serrana/noticia/2021/01/11/tragedia-na-serra-do-rj-completa-dez-anos-r-500-milhoes-ainda-devem-ser-investidos-em-obras.ghtml. Acesso em: 02 de Fevereiro de 2022.

Kobiyama, M., Mendonça, M., Moreno, D. A., Marcelino, I. P. d. O., Marcelino, E. V., Brazetti, L. L. P., Goerl, R. F., Molleri, M. G. S. F., e Rudorf, F. d. M. R. (2006). Prevenç ão de desastres naturais - - Conceitos Básicos. página 109.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference of Artificial Intelligence*, (March 2001).

Kohavi, R. e Provost, F. (1998). Glossary of terms. *Machine Learning*, 30(2-3):271–274.

Korup, O. e Stolle, A. (2014). Landslide prediction from machine learning. *Geology Today*, 30.

Kuhn, M. e Johnson, K. (2013). *Applied Predictive Modeling*. SpringerLink : Bücher. Springer New York.

Lauriano, C. (2010). Angra dos reis ainda se recupera das chuvas do réveillon de 2009. Disponível em: https://g1.globo.com/especiais/eleicoes-2010/noticia/2010/08/angra-dos-reis-ainda-se-recupera-das-chuvas-do-reveillon-de-2009.html. Acesso em: 02 de Fevereiro de 2022.

Liu, Z., Gilbert, G., Cepeda, J. M., Lysdahl, A. O. K., Piciullo, L., Hefre, H., e Lacasse, S. (2021). Modelling of shallow landslides with machine learning algorithms. *Geoscience Frontiers*, 12(1):385–393.

Logar, J., Turk, G., Marsden, P., e Ambrožič, T. (2017). Prediction of rainfall induced landslide movements by artificial neural networks. *Natural Hazards and Earth System Sciences Discussions*, 2017:1–18.

Loh, W. Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23.

Lucchese, L. V., de Oliveira, G. G., e Pedrollo, O. C. (2021a). Investigation of the influence of nonoccurrence sampling on landslide susceptibility assessment using Artificial Neural Networks. *Catena*, 198(November 2020):105067.

Lucchese, L. V., de Oliveira, G. G., e Pedrollo, O. C. (2021b). Mamdani fuzzy inference systems and artificial neural networks for landslide susceptibility mapping. *Catena*, 106(April 2021):2381–2405.

Ludermir, T. B. (2021). e Aprendizado de Máquina : I nteligência Artificial. 35(101):85–94.

Lumb, P. (1975). Slope failures in hong kong, quar-terly journal of engineering geology.

Macedo, E. S. d. (2019). Desastres naturais: mortes por deslizamento. *Revista Emergência (Novo Hamburgo)*, 121:28.

Maimon, O. e Rokach, L. (2010). *Data mining and knowledge discovery handbook*, volume 14. Springer.

Massad, F. (2010). *Obras de Terra: Curso Básico de Geotecnia*, volume 2ed. Oficina de textos, São Paulo.

Microsoft (2022). Lightgbm's documentation.

Montgomery, D. R. e Dietrich, W. E. (1994). A physically based model for the topographic control on shallow landsliding. *Water Resources Research*, 30(4):1153–1171.

Oliveira, F. R., Ouriques, J. M. d. A., e Correia, L. S. (2018). Percepção de risco a partir do programa Defesa Civil na Escola em Blumenau. *Territorium*, 25(II):5–18.

Pack, R. T., Tarboton, D. G., e N., G. C. (1998). The SINMAP Approach to Terrain Stability Mapping. *8th Congress of the International Association of Engineering Geology*, página 8.

Parizzi, M. G., Sebastião, C. S., Viana, C. d. S., Pflueger, M. d. C., Campos, L. d. C., Cajazeiro, J. M. D., Tomich, R. S., Guimarães, R. N., de Abreu, M. L., Sobreira, F. G., e dos Reis, R. (2010). Correlações entre chuvas e movimentos de massa no município de Belo Horizonte, MG. *Revista Geografias*, 6(2):49–68.

Pham, B. T., Prakash, I., Jaafari, A., e Bui, D. T. (2018a). Spatial Prediction of Rainfall-Induced Landslides Using Aggregating One-Dependence Estimators Classifier. *Journal of the Indian Society of Remote Sensing*, 46(9):1457–1470.

Pham, B. T., Prakash, I., e Tien Bui, D. (2018b). Spatial prediction of landslides using a hybrid machine learning approach based on Random Subspace and Classification and Regression Trees. *Geomorphology*, 303:256–270.

Prati, R. C., Batista, G. E. d. A. P. A., e Monard, M. C. (2008). Curvas roc para avaliação de classificadores. *IEEE Latin America Transactions*.

Qi, C. e Tang, X. (2018). Slope stability prediction using integrated metaheuristic and machine learning approaches: A comparative study. *Computers and Industrial Engineering*, 118:112–122.

Salaroli, I. S. (2003). *Movimentos de Massa no Município de Vitória-ES: inventário, caracterização e indicativos de um modelo comportamental*. Tese de Doutorado, Dissertação de Mestrado, Universidade Federal do Espírito Santo, Vitória.

Salles, R. e Amaral, C. (2013). Estudo da correlação entre chuvas e escorregamentos na região serrana do rio de janeiro. In *Conferência Brasileira de Encostas*, volume 6.

Santos, D. A. C. (2018). Análise de áreas suscetíveis a escorregamentos e da vulnerabilidade social em são marcos, salvador - bahia. Mestrado em geografia, Universidade Federal da Bahia, Salvador.

Soares, F. et al. (2015). Correlação entre movimentos de massa e pluviosidade nas encostas de joão pessoa/pb-brasil. *Geotecnia*, (133):51–62.

Souza, F. T. d. e Ebecken, N. F. (2012). A Data Based Model to Predict Landslide Induced by Rainfall in Rio de Janeiro City. *Geotechnical and Geological Engineering*, 30(1):85–94.

Souza, L. A. d. (2015). Salvador é esquartejada em novo processo de modernização. Disponível em: https://noticias.uol.com.br/opiniao/coluna/mobile/2015/07/09/salvador-e-esquartejada-em-novo-processo-de-modernizacao.htm. Acesso em: 04 de junho de 2021.

Tan, P., Steinbach, M., e Kumar, V. (2014). *Introduction to Data Mining*. Always learning. Pearson.

Tatizana, C., Ogura, A. T., Cerri, L. d. S., e Rocha, M. d. (1987). Análise de correlação entre chuvas e escorregamentos–serra do mar, município de cubatão. In *Congresso Brasileiro de Geologia de Engenharia*, volume 5, páginas 225–236.

Tehrani, F. S., Santinelli, G., e Herrera, M. (2019). A framework for predicting rainfall-induced landslides using machine learning methods. *17th European Conference on Soil Mechanics and Geotechnical Engineering, ECSMGE 2019 - Proceedings*.

Terzaghi, K. (1950). *Mecanismos de escorregamentos de terra. Trad. De Ernesto Pichler*. São Paulo.

Tien Bui, D., Tuan, T. A., Hoang, N. D., Thanh, N. Q., Nguyen, D. B., Van Liem, N., e Pradhan, B. (2016). Spatial prediction of rainfall-induced landslides for the Lao Cai area (Vietnam) using a hybrid intelligent approach of least squares support vector machines inference model and artificial bee colony optimization. *Landslides*, 14(2):447–458.

Tominaga, K., Santoro, J., e do. Amaral, R. (2009). *Desastres naturais: conhecer para prevenir*, volume 3ed. Instituto Geológico„ São Paulo.

Tran, H. (2019). Survey of machine learning and data mining techniques used in multimedia system.

Triola, M. F. (2018). *Elementary Statistics Technology Update*, volume 2011.

Tucci, C. (2020). *Hidrologia: ciência e aplicação*. ABRH. Coleção de Recursos Hídricos. Editora da Universidade, 9ª reimpressão 4ª edição.

Varnes, D. J. (1978). *Landslides Types and Processes*. Landslides and Engineering Pratice.

Vieira, R. (2004). Um olhar sobre a paisagem e o lugar como expressão do comportamento frente ao risco de deslizamento.

Wicander, R. e Monroe, J. S. (2009). *Fundamentos de Geologia*, volume volume. Cengage Learning, São Paulo.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. biometrics bulletin 1, 6 (1945), 80–83. *URL http://www. jstor. org/stable/3001968*.

Witten, I., Frank, E., e Hall, M. (2011a). *Data Mining: Practical Machine Learning Tools and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science.

Witten, I. H., Frank, E., e Hall, M. A. (2011b). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, Amsterdam, 3 edição.

Wolle, C. M. (1988). Análise dos escorregamentos translacionais numa região da serra do mar no contexto de uma classificação de mecanismos de instabilização de encostas. *São Paulo*.

Xavier, H. (1996). *Percepção geográfica dos deslizamentos de encostas em áreas de risco no município de Belo Horizonte, MG. 1996. 222p.* Tese de Doutorado, Tese (Doutorado em Geografia)–Instituto de Geociências, Universidade ....

Xiao, T., Yin, K., Yao, T., e Liu, S. (2019). Spatial prediction of landslide susceptibility using GIS-based statistical and machine learning models in Wanzhou County, Three Gorges Reservoir, China. *Acta Geochimica*, 38(5):654–669.

Yamamoto, J. K. e Landim, P. M. B. (2013). *Jorge Kazuo Yamamoto Paulo M. Barbosa Landim*.

Yi, Y., Zhang, Z., Zhang, W., e Xu, C. (2019). for Landslide Susceptibility Mapping. *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, páginas 9318–9321.

Zaki, M., Meira, W., e Meira, W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Zêzere, J., Rodrigues, M., e Ferreira, A. (2003). Recent landslide activity in relation to rainfall in the lisbon region (portugal). In *EGS-AGU-EUG Joint Assembly*, página 5506.

Zhu, X., Xu, Q., Tang, M., Nie, W., Ma, S., e Xu, Z. (2017). Comparison of two optimized machine learning models for predicting displacement of rainfall-induced landslide: A case study in Sichuan Province, China. *Engineering Geology*, 218:213–222.

Apêndice A

Cenários

A.1 Cenários com 16 dias de chuva acumulada

Tabela A.1: Cenário - Amostras negativas com 1 dia anterior ao escorregamento e previsão com 2 dias de antecedência.

Antecedência 2 dias	**Cenário 3**	Chuva até 9h do dia do escorregamento	
		Escorregamento	**Não escorregamento**
	Propriedades Geormoformétricas	G	G
	Caract. Geológicas e Geotécnicas	S	S
	Precipitação diária	[d-2, d-17]	[d-3, d-18]
	Precipitação acumulada	[d-2, d-17]	[d-3, d-18]
	Previsão de precipitação	[d*, d-1]	[d-1, d-2]
	Previsão de precipitação acum.	d-1	d-2
	Cenário 4	Chuva até 9h do dia posterior ao escorr.	
		Escorregamento	**Não escorregamento**
	Propriedades Geormoformétricas	G	G
	Caract. Geológicas e Geotécnicas	S	S
	Precipitação diária	[d-2, d-17]	[d-3, d-18]
	Precipitação acumulada	[d-2, d-17]	[d-3, d-18]
	Previsão de precipitação	[d+1**, d-1]	[d*, d-2]
	Previsão de precipitação acum.	[d*, d-1]	[d-1, d-2]

* Precipitação acumulada das 09:01h do dia d+1 até 09:00 do dia d;

** Precipitação acumulada das 09:01h do dia d+2 até 09:00 do dia d+1;

Tabela A.2: Cenário - Amostras negativas com 1 dia anterior ao escorregamento e previsão com 3 dias de antecedência.

	Cenário 5	Chuva até 9h do dia do escorregamento	
		Escorregamento	**Não escorregamento**
	Propriedades Geormoformétricas	G	G
	Caract. Geológicas e Geotécnicas	S	S
	Precipitação diária	[d-3, d-18]	[d-4, d-19]
	Precipitação acumulada	[d-3, d-18]	[d-4, d-19]
	Previsão de precipitação	[d*, d-2]	[d-1, d-3]
Antecedência 3 dias	Previsão de precipitação acum.	[d-1, d-2]	[d-2, d-3]
	Cenário 6	Chuva até 9h do dia posterior ao escorr.	
		Escorregamento	**Não escorregamento**
	Propriedades Geormoformétricas	G	G
	Caract. Geológicas e Geotécnicas	S	S
	Precipitação diária	[d-3, d-18]	[d-4, d-19]
	Precipitação acumulada	[d-3, d-18]	[d-4, d-19]
	Previsão de precipitação	[d+1**, d-2]	[d*, d-3]
	Previsão de Prec. acum	[d*, d-2]	[d-1, d-3]

* Precipitação acumulada das 09:01h do dia d+1 até 09:00 do dia d;
** Precipitação acumulada das 09:01h do dia d+2 até 09:00 do dia d+1;

Tabela A.3: Cenário - Amostras negativas geradas a partir da encosta mais próxima que escorregou em momento posterior e previsão com 2 dias de antecedência.

	Cenário 9	Chuva até 9h do dia do escorregamento	
		Escorregamento	**Não escorregamento**
	Propriedades Geormoformétricas	G	G_v
	Caract. Geológicas e Geotécnicas	S	S_v
	Precipitação diária	[d-2, d-17]	[d-2, d-17]
	Precipitação acumulada	[d-2, d-17]	[d-2, d-17]
	Previsão de precipitação	[d*, d-1]	[d*, d-1]
Antecedência 2 dias	Previsão de precipitação acum.	d-1	d-1
	Cenário 10	Chuva até 9h do dia posterior ao escorr.	
		Escorregamento	**Não escorregamento**
	Propriedades Geormoformétricas	G	G_v
	Caract. Geológicas e Geotécnicas	S	S_v
	Precipitação diária	[d-2, d-17]	[d-2, d-17]
	Precipitação acumulada	[d-2, d-17]	[d-2, d-17]
	Previsão de precipitação	[d+1**, d-1]	[d+1**, d-1]
	Previsão de Prec. acum.	[d*, d-1]	[d*, d-1]

* Precipitação acumulada das 09:01h do dia d+1 até 09:00 do dia d;
** Precipitação acumulada das 09:01h do dia d+2 até 09:00 do dia d+1;

Tabela A.4: Cenário - Amostras negativas geradas a partir da encosta mais próxima que escorregou em momento posterior e previsão com 3 dias de antecedência.

	Cenário 9	Chuva até 9h do dia do escorregamento	
		Escorregamento	**Não escorregamento**
	Propriedades Geormoformétricas	G	G_v
	Caract. Geológicas e Geotécnicas	S	S_v
	Precipitação diária	[d-3, d-18]	[d-3, d-18]
	Precipitação acumulada	[d-3, d-18]	[d-3, d-18]
	Previsão de precipitação	[d*, d-2]	[d*, d-3]
Antecedência	Previsão de precipitação acum.	[d-1, d-2]	[d-1, d-2]
3 dias			
	Cenário 10	Chuva até 9h do dia posterior ao escorr.	
		Escorregamento	**Não escorregamento**
	Propriedades Geormoformétricas	G	G_v
	Caract. Geológicas e Geotécnicas	S	S_v
	Precipitação diária	[d-3, d-18]	[d-3, d-18]
	Precipitação acumulada	[d-3, d-18]	[d-3, d-18]
	Previsão de precipitação	[d+1**, d-2]	[d+1**, d-2]
	Previsão de Prec. acum.	[d*, d-2]	[d*, d-2]

* Precipitação acumulada das 09:01h do dia d+1 até 09:00 do dia d;
** Precipitação acumulada das 09:01h do dia d+2 até 09:00 do dia d+1;

A.2 Cenários com 8 dias de chuva acumulada

Tabela A.5: Cenário - Cenário 13 - amostras negativas com 1 dia anterior ao escorregamento e previsão com 1 dia de antecedência e Cenário 14 - amostras negativas a partir da encosta mais próxima que escorregou em momento posterior ao escorregamento e previsão com 1 dia de antecedência.

	Cenário 13	Chuva até 9h do dia do escorregamento	
		Escorregamento	**Não escorregamento**
	Propriedades Geormoformétricas	G	G
	Caract. Geológicas e Geotécnicas	S	S
	Precipitação diária	[d-1, d-8]	[d-2, d-9]
	Precipitação acumulada	[d-1, d-8]	[d-2, d-9]
	Previsão de precipitação	d*	d-1
Antecedência			
1 dia	**Cenário 14**	Chuva até 9h do dia do escorregamento	
		Escorregamento	**Não escorregamento**
	Propriedades Geormoformétricas	G	G_v
	Caract. Geológicas e Geotécnicas	S	S_v
	Precipitação diária	[d-1, d-8]	[d-1, d-8]
	Precipitação acumulada	[d-1, d-8]	[d-1, d-8]
	Previsão de precipitação	d*	d*

* Precipitação acumulada das 09:01h do dia d+1 até 09:00 do dia d;
** Precipitação acumulada das 09:01h do dia d+2 até 09:00 do dia d+1;

A.3 Cenários com 4 dias de chuva acumulada

Tabela A.6: Cenário - Cenário 15 - amostras negativas com 1 dia anterior ao escorregamento e previsão com 1 dia de antecedência e Cenário 16 - amostras negativas a partir da encosta mais próxima que escorregou em momento posterior ao escorregamento e previsão com 1 dia de antecedência.

	Cenário 15	Chuva até 9h do dia do escorregamento	
		Escorregamento	**Não escorregamento**
	Propriedades Geormoformétricas	G	G
	Caract. Geológicas e Geotécnicas	S	S
	Precipitação diária	[d-1, d-4]	[d-2, d-5]
	Precipitação acumulada	[d-1, d-4]	[d-2, d-5]
	Previsão de precipitação	d*	d-1
Antecedência 1 dia			
	Cenário 16	Chuva até 9h do dia do escorregamento	
		Escorregamento	**Não escorregamento**
	Propriedades Geormoformétricas	G	G_v
	Caract. Geológicas e Geotécnicas	S	S_v
	Precipitação diária	[d-1, d-4]	[d-1, d-4]
	Precipitação acumulada	[d-1, d-4]	[d-1, d-4]
	Previsão de precipitação	d*	d*

* Precipitação acumulada das 09:01h do dia d+1 até 09:00 do dia d;

** Precipitação acumulada das 09:01h do dia d+2 até 09:00 do dia d+1;