




**Programa de Pós-Graduação em Instrumentação, Controle e Automação de Processos de Mineração (PROFICAM)**

**Escola de Minas, Universidade Federal de Ouro Preto (UFOP)**

**Associação Instituto Tecnológico Vale (ITV)**

**Dissertação**

# APRENDIZADO DE MÁQUINA APLICADO EM PREVISÃO DE CURTO PRAZO DE VALORES DE INDICADORES DE NÍVEL DE ÁGUA

**Luiz Frederico de Freitas Kümmel**

**Ouro Preto**

**Minas Gerais, Brasil**

**2021**

**Luiz Frederico de Freitas Kümmel**

# APRENDIZADO DE MÁQUINA APLICADO EM PREVISÃO DE CURTO PRAZO DE VALORES DE INDICADORES DE NÍVEL DE ÁGUA

Dissertação apresentada ao Programa de Pós-Graduação em Instrumentação, Controle e Automação de Processos de Mineração da Universidade Federal de Ouro Preto e do Instituto Tecnológico Vale, como parte dos requisitos para obtenção do título de Mestre em Engenharia de Controle e Automação.

Orientador: Gustavo Pessin, D.Sc.
Coorientador: Vidal Félix Navarro Torres, Ph.D.
Coorientador: Jodelson Sabino, D.Sc.

Ouro Preto
2021

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

K95a Kummel, Luiz Frederico de Freitas.
Aprendizado de máquina aplicado em previsão de curto prazo de valores de indicadores de nível de água. [manuscrito] / Luiz Frederico de Freitas Kummel. - 2021.
72 f.

Orientador: Prof. Dr. Gustavo Pessin.
Coorientadores: Dr. Jodelson Aguilar Sabino, Prof. Dr. Vidal Félix Navarro Torres.
Dissertação (Mestrado Profissional). Universidade Federal de Ouro Preto. Programa de Mestrado Profissional em Instrumentação, Controle e Automação de Processos de Mineração. Programa de Pós-Graduação em Instrumentação, Controle e Automação de Processos de Mineração.
Área de Concentração: Engenharia de Controle e Automação de Processos Minerais.

1. Barragens de rejeitos. 2. Aprendizado de máquina. 3. Indicadores de nível. I. Pessin, Gustavo. II. Sabino, Jodelson Aguilar. III. Torres, Vidal Félix Navarro. IV. Universidade Federal de Ouro Preto. V. Título.
CDU 681.5:624.136

Bibliotecário(a) Responsável: Sione Galvão Rodrigues - CRB6 / 2526

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
ESCOLA DE MINAS
PROGR. POS GRAD. PROF. INST. CONT. E AUT.
PROCESSOS DE MIN.

**FOLHA DE APROVAÇÃO**

**Luiz Frederico de Freitas Kümmel**

**Aprendizado de Máquina Aplicado em Previsão de Curto Prazo de Valores de Indicadores de Nível de Água**

Dissertação apresentada ao Programa de Pós-Graduação em Instrumentação, Controle e Automação de Processos de Mineração (PROFICAM), Convênio Universidade Federal de Ouro Preto/Associação Instituto Tecnológico Vale - UFOP/ITV, como requisito parcial para obtenção do título de Mestre em Engenharia de Controle e Automação na área de concentração em Instrumentação, Controle e Automação de Processos de Mineração.

Aprovada em 08 de setembro de 2021

Membros da banca

Doutor - Gustavo Pessin - Orientador - Instituto Tecnológico Vale
Doutor - Jodelson Aguilar Sabino - Coorientador - Vale
Doutor - Juan Manuel Girao Sotomayor - Instituto Tecnológico Vale
Ph.D - Renato Hidaka - Universidade Federal do Pará

Gustavo Pessin, orientador do trabalho, aprovou a versão final e autorizou seu depósito no Repositório Institucional da UFOP em 29/11/2021

Documento assinado eletronicamente por **Bruno Nazário Coelho**, **COORDENADOR(A) DE CURSO DE PÓS-GRAD EM INSTRUMENTAÇÃO CONTROLE E AUTOMAÇÃO DE PROCESSOS DE MINERAÇÃO**, em 01/12/2021, às 23:55, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0252809** e o código CRC **77779C13**.



R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35400-000
Telefone: - www.ufop.br

## Agradecimentos

Agradeço primeiramente a Deus pela vida e o mundo de oportunidades que o mundo nos traz. Agradeço a minha mãe Maria Antônia e meu pai Luiz Carlos que me proporcionaram uma educação admirável me ajudando a ser um ser humano com caráter ilibado. Um agradecimento especial a minha esposa Suelen que sempre me apoiou nos meus sonhos. Agradeço também a Vale e ITV por me proporcionarem oportunidades tão sublimes em meus crescimentos profissionais e acadêmicos. E por fim meu agradecimento ao Gustavo Pessin que sempre me orientou de forma magistral abrindo minha mente de forma a pensar fora da caixa!

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil (CAPES), Código de Financiamento 001; do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); da Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG); e da Vale SA.

*“O homem erudito é um descobridor de fatos que já existem - mas o homem sábio é um criador de valores que não existem e que ele faz existir”.*

*(Albert Einstein).*

## Resumo

Resumo da Dissertação apresentada ao Programa de Pós-Graduação em Instrumentação, Controle e Automação de Processos de Mineração como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

APRENDIZADO DE MÁQUINA EM PREVISÃO DE CURTO PRAZO DE VALORES DE INDICADORES DE NÍVEL DE ÁGUA

Luiz Frederico de Freitas Kümmel

Setembro/2021

Orientadores: Gustavo Pessin

Vidal Félix Navarro Torres

Jodelson Sabino

A estabilidade e solidez de barragens de rejeito para resíduos de atividades industriais de mineração é de importância primordial para a segurança da sociedade e meio ambiente localizado a sua jusante. Para assegurar as essenciais exigências de segurança e exposição ao risco das barragens ao longo da sua vida útil, devem ser implementadas ações mitigatórias de prevenção e controle dessas condições, nesse intuito esse trabalho visa aplicar métodos de *Machine Learning*, para prever o comportamento dos indicadores de nível de água associados a carta de risco. Os algoritmos de *machine learning* mostraram elevadas taxas de acerto para predição, sendo que a combinação de métodos de classificação e regressão permitiu aumentar ainda mais a qualidade de resposta do sistema proposto. .

**Palavras-chave**: Barragens de Rejeito, Indicador de Nível De água, *Machine Learning*.

**Macrotema:** Mina; **Linha de Pesquisa:** Tecnologias da Informação, Comunicação e Automação Industrial; **Tema:** Saúde e Segurança; **Área Relacionada da Vale:** Geotecnia.

## Abstract

Abstract of Dissertation presented to the Graduate Program on Instrumentation, Control and Automation of Mining Process as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

MACHINE LEARNING APPLIED IN SHORT OF WATER LEVEL INDICATOR VALUE

Luiz Frederico de Freitas Kümmel

September/2021

Advisors: Gustavo Pessin

Vidal Félix Navarro Torres

Jodelson Sabino

The stability and solidity of tailings dams for residues from industrial mining activities is of paramount importance for the safety of society and the environment located downstream. To ensure the essential safety and risk exposure requirements of dams throughout their useful life, mitigation actions must be implemented to prevent and control these conditions. To this end, this work aims to apply Machine Learning methods to predict the behavior of water level associated with the risk chart. Machine learning algorithms showed high success rates for prediction, and the combination of classification and regression methods allowed to further increase the response quality of the proposed system.



**Keywords**: Tailings Dams, Water Level Indicator, Machine Learning.

**Macrotheme:** Mine; **Research Line:** Information Technology; Communication; e Industrial Automation; **Theme:** Health and Safety; **Related Area of Vale:** Geotechnics.

## Lista de Figuras

## Lista de Tabelas

## Lista de Siglas e Abreviaturas

ARIMA: *Auto Regressive Integrated Moving Avarage*.

AUC: *Area under the ROC curve.*

FS: Fator de Segurança.

INA: Indicador de Nível De água.

ITV: Instituto Tecnológico Vale.

LSTM: *Long Short-Term Memory*.

MAE: *Mean Absolute Error.*

MDPI: *Multidisciplinary Digital Publishing Institute.*

ML: *Machine Learning*.

MLP: *Multi-Layer Perceptron*.

MSE: *Mean Squared Error*.

PFI: *Permutation Feature Importance*.

PSB: Plano de Segurança da Barragem.

PVC: Poli Cloreto de Vinila.

RMSE: *Root Mean Square Error*.

ROC: *Receiver Operating Characteristic*.

SHMS: *Slope Health Monitoring System*.

## Sumário

# 1. Introdução

## 1.1. Contextualização

Na indústria de mineração, enormes volumes e massas de materiais minerais passam pelos processos de desmonte, carreamento e transporte.

A Figura 1 apresenta de forma macro um processo típico de mineração, conforme IBRAM (2016), a etapa (1) é a execução do desmonte na mina, onde os materiais estéreis são empilhados nas pilhas de estéril (2) e o minério é encaminhado para a usina de concentração (3) onde existem os processos industriais físico-químicos para que haja a separação do rejeito do minério concentrado, o material que tem valor econômico é escoado por comboios de trens e vagões através de linha férrea (4) para o cliente final. Os rejeitos oriundos dos processos de beneficiamento realizados dentro da usina de concentração, são encaminhados para serem armazenados na barragem de rejeitos (5), por fim, há a recuperação de água (6) do reservatório da barragem, para que a mesma seja reaproveitada dentro da usina de concentração para processos que necessitam da mesma.

De acordo com IPEA (2012) na atividade de mineração existem a presença de dois tipos principais de resíduos sólidos: os estéreis e os rejeitos. Os estéreis são os materiais originados de escavações gerados pelas atividades de lavra na retirada do material estéril da mina, eles não possuem valor econômico atrativo e ficam dispostos nas pilhas de estéril. Os rejeitos são resíduos resultantes dos processos de beneficiamento de cominuição mecânico ou químico, que separam o minério concentrado do rejeito.

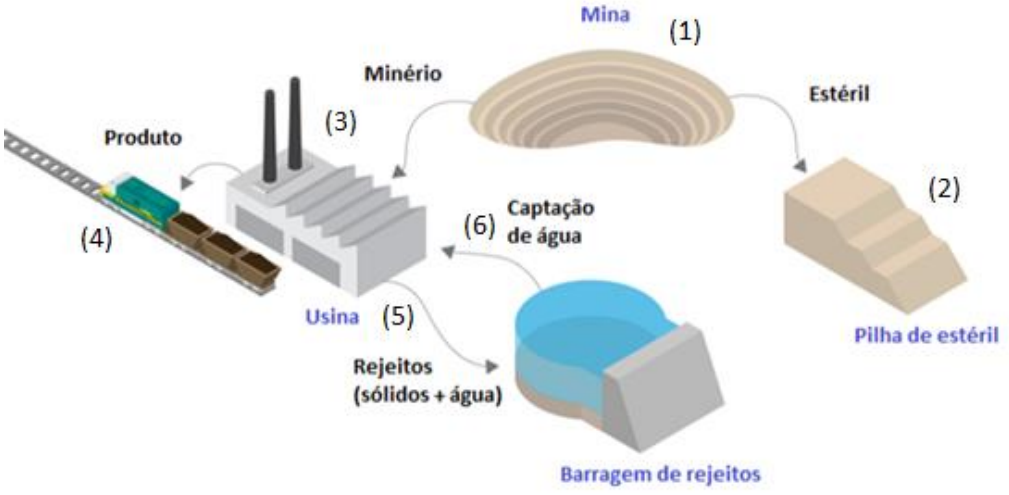


Figura 1: Visão macro de uma operação de mineração. Adaptado, Fonte: ITV (2020).

De acordo com IBRAM (2016), a disposição de rejeitos em reservatórios criados por diques ou barragens é o método utilizado na área de mineração para armazenamento dos mesmos, estas barragens ou diques, podem ser de solo natural ou construídos com os próprios rejeitos, sendo classificadas, como barragens de contenção alteadas com rejeitos. Os rejeitos são transportados para a área de disposição com um alto teor de água e com 10% a 25% de sólidos, entretanto, estes rejeitos devem ser criteriosamente armazenados para segurança de pessoas, meio ambiente e instalações industriais.

No Boletim 68 ICOLD (1989) é destacado que a instrumentação geotécnica realiza medições de grandezas físicas, que de fato, contribuem para realização de análises de estabilidade das estruturas geotécnicas. Estas medições fornecem uma base de dados para entender melhor o comportamento dinâmico e estabilidade das barragens de rejeitos.

Dentro desse contexto a disciplina de ciência de dados pode desempenhar um papel importante provendo informações preditivas para equipes de sustentabilidade de barragens para tomadas de decisões. De posse de informações fidedignas fornecidas pela instrumentação geotécnica é possível desenvolver algoritmos de regressão e classificação baseados em diversas técnicas de *machine learning* (ML) para prever os valores dos instrumentos de monitoramento geotécnico que estão correlacionados com a estabilidade de uma barragem de rejeitos e através da comparação dessas previsões com os valores de segurança determinados na carta de risco, prover informações às equipes técnicas de monitoramento que as possibilitem agir de forma preventiva para garantir uma operação segura, eficiente e sustentável de estruturas geotécnicas.

A carta de risco é o documento no qual são apresentados os níveis de controle da instrumentação geotécnica de uma barragem para sua estabilidade. É uma ferramenta usualmente adotada na gestão de segurança de barragens. Além disso, destaca-se que, a definição dos níveis de controle para a instrumentação de barragens é um requisito legal, conforme estipulado na Portaria do DNPM (atual ANM) n° 70.389 de 2017, devendo ser apresentada no Plano de Segurança de Barragens (PSB).

De acordo com MI (2002), o estabelecimento dos níveis de controle ocorre a partir da realização das análises de estabilidade para diferentes condições de níveis piezométricos e de lençóis freáticos, de modo a obter Fatores de Segurança de referência (FS = 1,5; FS = 1,3 e FS = 1,1), onde FS é o coeficiente de segurança ao deslizamento. Dessa forma, para cada

instrumento geotécnico é estabelecido o valor da leitura associada aos níveis: normal, atenção, alerta e emergência inseridos na carta de riscos da barragem de rejeitos.

## 1.2. Motivação e justificativa

Uma das maiores preocupações e pontos de atenção relacionados a segurança de barragens de rejeito é a estabilidade do talude de jusante da mesma, pois, condições inseguras podem levar a um deslizamento que ameaça a comunidade, meio ambiente e instalações industriais localizadas a jusante da mesma.

Conforme o Instituto Minere (2016), um fator determinante para a estabilidade de uma barragem é a localização da linha freática dentro do aterro, uma linha freática pode diminuir consideravelmente a capacidade de resistir a deslizamentos, o que pode ser observado por percolações que ocorrem na estrutura.

Já existem softwares de mercado, como por exemplo, o *Slope Health Monitoring System* (SHMS) de fornecimento da Intelltech (2021), Rocscience de fornecimento da Rocscience (2021) e o Qgis de fornecimento da QGIS (2021), que processam dados de instrumentação geotécnica e utilizam valores de alertas determinísticos, que, se extrapolados geram alarmes automáticos para acionar as equipes de sustentabilidade de estruturas geotécnicas para a atuação mitigatória de anomalias, porém, esses programas são de características de atuação de alarmes reativos e não preventivos.

Neste âmbito, percebe-se uma oportunidade de empregar métodos baseados em *machine learning* para realizar predições dos valores da instrumentação geotécnica que está associada a estabilidade da barragem de rejeitos, tornando o processo de atuação de alertas preditivo e não reativo, auxiliando assim as equipes de sustentabilidade de barragens, tomar decisões de manutenção de caráter preventivo.

A principal motivação deste trabalho é criar modelos preditivos de regressão e classificação através de técnicas de modelagem baseadas em ML para previsão de valores futuros do instrumento indicador de nível de água (INA), tal instrumento, determina de maneira exata a posição da linha freática no interior de um maciço de uma barragem, tal grandeza é diretamente correlacionada à estabilidade da estrutura. Desse modo, serão criados métodos de aprendizado de máquina através de algoritmos baseados em *random forest, AdaBoost,* regressão

linear, redes neurais, árvores de decisão e *Stochastic Gradient Descent* (SGD), para avaliar de forma preditiva, a detecção de cargas hidráulicas que possam afetar o sistema de drenagem de uma estrutura geotécnica e por consequência sua estabilidade.

O intuito dessa pesquisa é analisar os resultados que serão gerados através dos métodos de ML e apontar qual técnica se adapta melhor ao problema apresentado. Posteriormente, aplicar um protótipo em produção, de forma que o mesmo auxilie as equipes de sustentabilidade de barragens de rejeito nas tomadas de decisões de forma preventiva, para garantir a estabilidade e segurança de estruturas geotécnicas.

## 1.3. Questões de pesquisa

De acordo com o Instituto Minere (2016) as medições das grandezas geotécnicas do instrumento indicador de nível de água (INA) são de grande importância para a averiguação da capacidade da drenagem interna e dos sistemas de vedação de uma barragem de rejeitos. A partir da análise desses dados, pode-se certificar com grande assertividade a rede de fluxo e estabilidade de uma barragem de rejeitos.

Algoritmos preditivos que proporcionem medições futuras do instrumento INA é de grande importância para as equipes de sustentabilidade de barragens de rejeito, para que as mesmas possam a partir dessas previsões diligenciar ações estruturantes de campo de características mitigatórias para o aumento de segurança e estabilidade das estruturas geotécnicas.

Serão utilizados nessa pesquisa os valores históricos de dois indicadores de nível de água e um pluviômetro para as regressões e classificações baseadas em ML, nessa avaliação, as seguintes questões devem ser respondidas:

- Qual a capacidade de predição usando apenas histórico para indicadores de nível de água com comportamentos suaves e abruptos?
- Qual o efeito das janelas de memória para os sistemas preditivos usando apenas dados históricos?
- Existe como avaliar sistema de classificação para o caso do indicador de nível de água fora do limite de leitura do sensor?
- Quais resultados o sistema retorna com inclusão de dados de pluviometria?

Para essa pesquisa serão utilizadas as séries históricas da instrumentação instalada na barragem Capitão do Mato de propriedade da Vale S/A.

## 1.4. Objetivos gerais e específicos

### a) Objetivos gerais

O objetivo geral desse trabalho é realizar previsões de futuras medições de indicadores de nível de água, mediante o aprendizado de máquina, com a finalidade de comparar com os valores da carta de riscos para uma atuação preventiva de ações a serem diligenciadas na estrutura no sentido de garantir sua segurança e estabilidade.

### b) Objetivos específicos

- Propor modelos de regressão e classificação em cascata para prever valores futuros de indicadores de nível de água associados ou não as previsões pluviométricas, através de algoritmos fundamentados nas técnicas de *machine learning*.
- Prever o comportamento dos instrumentos antecipadamente e caso necessário acionar equipes de manutenção e sustentabilidade de barragens para intervenções antecipadas, garantindo alta segurança, estabilidade e confiabilidade da operação da barragem de Capitão do Mato.

## 1.5. Estrutura da dissertação

Essa dissertação apresenta-se dividida em seis capítulos. No Capítulo 1 são expostas a contextualização, motivação, justificativa do trabalho, as questões de pesquisa, os objetivos gerais e específicos a serem abordados no decorrer da pesquisa.

O Capítulo 2 é uma revisão bibliográfica que traz os elementos componentes de uma barragem de rejeitos e fundamentos de instrumentação geotécnica, nesse capítulo também são abordados os trabalhos correlacionados a esse estudo.

No Capítulo 3 são relatados os materiais e métodos utilizados no estudo, onde são apresentadas as fontes de extração dos dados dos instrumentos geotécnicos, o tratamento dos dados, apresentação dos gráficos de séries temporais dos indicadores de nível de água e concentração pluviométrica e os métodos de ML aplicados nessa pesquisa.

O Capítulo 4 apresenta os resultados obtidos utilizando as técnicas de ML, esse capítulo traz um planejamento da implementação de um projeto piloto.

O Capítulo 5 é a parte das conclusões do estudo que foram de interesse desse estudo, nesse capítulo todas as questões de pesquisa são respondidas.

O Capítulo 6 discute sobre as recomendações para trabalhos futuros que possam ser desenvolvidos para aprimorar as técnicas de regressão e classificação.

# 2. Revisão bibliográfica e tecnológica

É de suma importância descrever os elementos das estruturas associadas a uma barragem de rejeitos de mineração no sentido de explanar seus principais componentes, será feita também uma abordagem dos métodos de aprendizado de máquina, instrumentação geotécnica e trabalhos relacionados.

## 2.1. Estrutura de uma barragem de rejeitos

Uma barragem de rejeitos contempla diversas estruturas em sua composição, na Figura 2 são apresentadas as principais estruturas bem como as definições de cada uma delas.

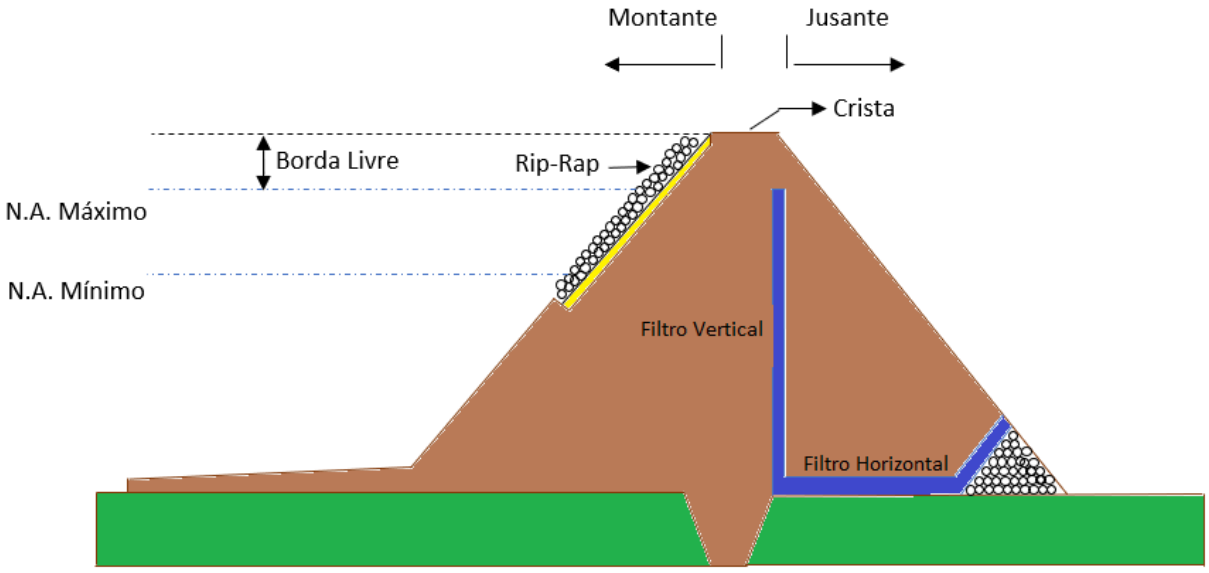


Figura 2: Estrutura de uma barragem de rejeitos. Fonte: Autor (2021).

As definições de componentes de barragem de rejeito, conforme MI (2002), são descritas subsequentemente.

Borda livre: A borda livre é a distância vertical entre a crista da barragem e o nível das águas do reservatório cujo objetivo é prover a segurança contra o galgamento, evitando dessa maneira danos e erosão no talude a jusante.

Crista: A crista é a parte superior da barragem. Sua largura é concebida através do tráfego de veículos e pessoas sobre ela. A altura da barragem deve ser no mínimo equivalente ao nível máximo da água adicionando a borda livre.

Filtro Vertical: Dispositivo interno da barragem responsável pela drenagem de água na posição vertical.

Filtro Horizontal: Dispositivo interno da barragem responsável pela drenagem de água na posição horizontal.

Rip-rap: É utilizado para proteger superficialmente o talude, é usado para estancar erosões.

## 2.2. Técnicas de aprendizado de máquina

O advento do aprendizado de máquina está realizando melhorias significativas na ciência de dados para a criação de modelos preditivos, sendo utilizados em diversos campos de atuação da sociedade.

Criminisi e Shotton (2013) descrevem que as técnicas de ML são métodos de análises de dados que automatizam a construção de modelos analíticos sendo fundamentados na ideia de que os sistemas podem aprender com dados históricos de um problema qualquer, de forma que, eles identificam padrões e são capazes de tomar decisões.

Conforme Aria et al. (2021), os algoritmos de aprendizado de máquina normalmente são usados em problemas onde não se tenha uma equação de solução, o que geralmente é feito, é obter uma massa de dados históricos que representam um problema a ser solucionado, onde esses dados são utilizados para serem treinados e encontrar uma possível solução, salientando que os dados devem ser de qualidade e representativos do problema a ser tratado, para esse estudo serão utilizadas seis técnicas de ML, sendo que a que obtiver o melhor resultado em regressão baseada na métrica de RMSE será desdobrada para avaliações de classificação e regressão em cascata.

## 2.3. Fundamentos de instrumentação geotécnica

A instrumentação geotécnica é de enorme importância nesse trabalho, ela que subsidiará todas as informações necessárias para que se possam realizar análises estatísticas e desenvolvimento de algoritmos baseados em aprendizado de máquina de forma a prever valores futuros dos indicadores de nível de água.

Cruz (1996) comenta que a instrumentação geotécnica auxilia no processo de verificação da performance e tendência de comportamento de uma estrutura geotécnica, sendo ela fundamental para estudos analíticos.

Cerqueira (2017) comenta que vários são os instrumentos para monitoramento de grandezas geotécnicas em uma barragem, sendo eles e não se limitando a: Poropressão, nível de água do reservatório, nível de lençóis freáticos, vazão, inclinação, percolação, sismos entre outros. De posse desses dados as mineradoras conseguem operar barragens de forma a otimizar o seu aproveitamento e adequando o projeto da barragem às condições de extração dos minerais.

Fusaro (2007) salienta que, as boas práticas de controle de segurança de uma barragem determina que existam inspeções visuais periódicas do maciço e estruturas auxiliares e monitoramento por meio de instrumentação geotécnica, nesse quesito é necessário, coletar, verificar e validar os dados advindos da instrumentação e interpretar os dados de forma analítica.

Conforme Silveira (2006), a instrumentação geotécnica tem que ter uma funcionalidade específica para que possa ser instalada em estruturas geotécnicas, a definição de qual grandeza geotécnica deve ser monitorada, o tipo de instrumento e a posição do instrumento na barragem é de responsabilidade do geotécnico responsável pela estrutura. Após instalada a instrumentação e sistema, os mesmos devem ser capazes de realizar a aquisição, transformação e registro dos dados geotécnicos.

De acordo com Soares (2010), o objetivo da instrumentação geotécnica é de garantir a segurança ambiental de pessoas e infraestrutura associada ao empreendimento e deve ter como direcionamento todas as fases do ciclo de vida da estrutura, que são: implantação, operação, paralisação e desativação.

## 2.4. Instrumento para medição de nível de lençol freático (indicador de nível de água, INA)

De acordo com Cerqueira (2017), a medição com precisão do posicionamento da linha do lençol freático que percola pelo maciço de uma barragem é essencial para uma observação detalhada no sentido de avaliar e interpretar a estabilidade de uma estrutura geotécnica. O princípio de funcionamento do indicador de nível de água é baseado no contato do elemento sensor diretamente da linha de fluxo de água que percola pelo maciço da barragem, no sentido de medir a cota de água de superfície, a Figura 3 apresenta as linhas de fluxo em uma barragem.

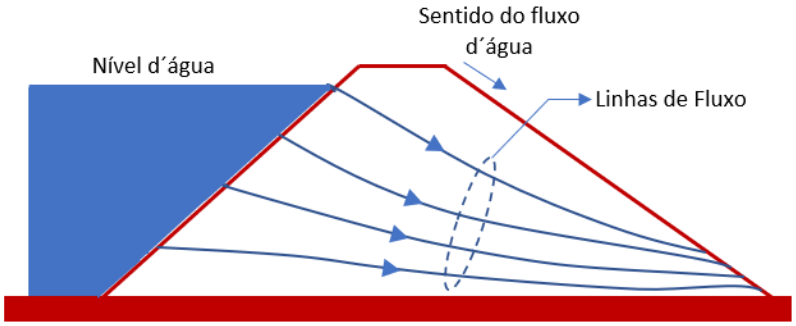


Figura 3: Barragem e suas linhas de fluxo. Fonte: Autor (2021).

A instalação dos indicadores de nível de água é realizada através de um tubo de PVC perfurado que é inserido no furo de sondagem, envolto por material filtrante (geotêxtil) e drenante (areia). É aplicada camada selante para vedar o espaço anular superior entre o tubo e o furo, a Figura 4 apresenta os elementos descritos.

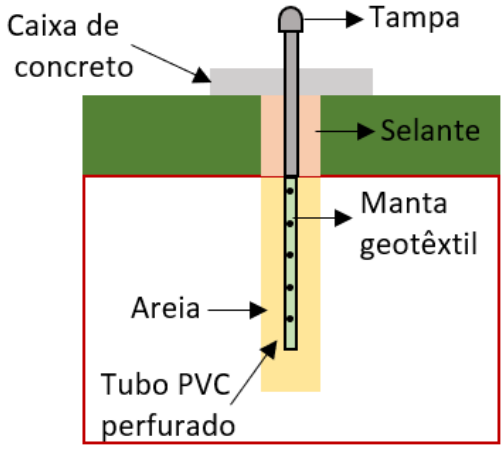


Figura 4: Elementos de instalação de um indicador de nível de água fluxo. Fonte: Autor (2021).

Pelo método de instalação do indicador de nível de água no maciço de uma barragem, este instrumento realiza medições do nível médio da água do solo, já que possibilita a comunicação vertical entre dois ou mais lençóis freáticos, podendo ser até num mesmo, quando existem fluxos descendentes ou ascendentes.

Dentro do tubo de PVC o nível de água sofre mudanças da mesma maneira da média de variação do nível de água das camadas saturadas em que a prospecção de sondagem interceptou.

A medição da variação do nível de água é realizada através de um ponto de referência (cota de topo, por exemplo), essa variação pode ser estabelecida através de um dispositivo

denominado por medidor de nível de água. Esse medidor é composto por uma trena milimetrada acoplada em um apito (pio) ou em um eletrodo sensor elétrico que envia sinais sonoros em contato com a água. A Figura 5 apresenta esse medidor.


Figura 5: Medidor de nível de água. Fonte: Geokon (2020).

## 2.5. Pluviômetro

Conforme Bedoya et al. (2017), o pluviômetro é um equipamento de meteorologia que é utilizado para coletar e medir a quantidade de líquidos ou sólidos precipitados durante um determinado tempo e local, a sua unidade de medida é de milímetros lineares por metro quadrado. Este instrumento traz como resultado o índice pluviométrico, que é o somatório da precipitação em um determinado local durante um tempo estabelecido. O regime pluviométrico consiste na distribuição das chuvas durante os doze meses de um ano, ele é representado através de climogramas por colunas mensais, através de sua análise é possível caracterizar o regime e com isso o índice pluviométrico.

Figura 6: Pluviômetro. Fonte: Sigma Sensors (2020).

## 2.6. Trabalhos relacionados

No sentido de complementar o estudo, foram realizadas pesquisas em diversos artigos de trabalhos correlatos com o objetivo de enriquecer ainda mais essa dissertação. Os artigos pesquisados foram adquiridos através dos repositórios *Sensors* e *Science Direct*, este são repositórios operacionalizados pela MDPI (*Multidisciplinary Digital Publishing Institute*) e editora *Elsevier* respectivamente, esses artigos foram publicados entre os anos de 2006 até 2020.

De acordo com Furquim et al. (2018), os modelos de previsões auxiliam no entendimento e comportamento das variáveis de um estudo de caso, tanto em situações regulares ou adversas ajudando na identificação de padrões de comportamento, e comentado também que deve se atentar o impacto que cada variável de um problema afeta nos resultados estimados e que diferentes janelas de tamanho na entrada dos modelos impactam na precisão de uma previsão. O uso de técnicas de seleção de atributos ajuda no alcance de melhores características no sentido de melhorar acurácia. A utilização da *Permutation Feature Importance* (PFI), que é uma técnica que seleciona melhores janelas de tempos para variáveis climáticas traz muitos ganhos nas predições, pois ela, orienta o processo de remoção de recursos desnecessários ou perturbadores no sentido de melhorar o RMSE. Por fim é importante comparar os resultados estatisticamente

através dos testes de normalidade de Shapiro-Wilk no sentido de certificar a distribuição dos resultados e após aplicar o teste de duas amostras para determinar a semelhança dos resultados.

Na visão de Furquim et al. (2016), desastres naturais são problemas que afetam as vidas e perdas financeiras de forma global, o monitoramento de ambientes naturais é muito desafiador pelo fato de suas características intrínsecas e é essencial fomentar diversos estudos para lidar com os dados históricos para a melhor modelagem no sentido de possibilitar a previsão em desastres que possam estar iminentes. De posse de dados históricos é possível estudar, modelar e investigar questões relativas à previsão. Nesse contexto a teoria do caos possibilita interpretar séries temporais de forma não trivial, ou seja, no caso de se tratar dados de maneira variada pode melhorar resultados de precisão das técnicas de ML. Através de uma abordagem do teorema da incorporação uma série temporal pode ser reconstruída em vetores, sendo que eles representam relações interdependentes entre observações aumentando a precisão da previsão, após a aplicação desse teorema o conjunto de técnicas de ML podem ser implementadas para criação da previsibilidade.

É sustentado por Uélisson et al. (2019), que muitos dos problemas de predição não resolvidos ocorrem por falta de análises mais aprofundadas de dados coletados e que diversas observações se limitam a executar atividades reativas, pois, de fato elas só são realizadas quando um nível determinístico de alerta for alcançado, que embora simples pode trazer problemas uma vez que a ação é realizada só após um limite determinado for ultrapassado. A utilização da técnica *Auto Regressive Integrated Moving Avarage* (ARIMA) é uma estratégia para utilizar em séries temporais cujo objetivo não está na elaboração de um modelo ou equações, mas sim, em avaliar as propriedades probabilísticas dessas séries, sendo essa técnica adequada para descrever séries não estacionárias, ou seja, séries que não possuem média constante no período avaliado.

De acordo com Ueyama et al. (2017) diversos sistemas adaptativos podem ser utilizados em eventos de acidentes naturais e serem tratados como sistemas de missão crítica, pois, envolvem riscos de perda de vidas.

Em outra abordagem Furquim et al. (2018) afirma que existiram grandes avanços promissores nas tecnologias de previsão de desastres naturais no sentido de complementar o monitoramento do meio ambiente, pois, as forças da natureza são brutais e imprevisíveis podendo ocasionar danos materiais e ceifar vidas humanas. No gerenciamento de

previsibilidade de acidentes naturais, as informações que são coletadas pelos sensores de campo devem ser analisadas juntamente com dados disponíveis na internet, como por exemplo, a previsão meteorológica. Utilizando métodos baseados em ML para a previsão de desastres naturais a tomada de decisões fica mais rápida devida à velocidade e precisão que essas técnicas trazem para auxílio em diversos tipos de problemas, sendo indispensável os artifícios tecnológicos para tolerância a falhas.

Para Vieira et al. (2020), um fator determinante para sistemas de previsão é efetivamente escolher as melhores características com a finalidade de aprimorar os modelos que estimam desastres naturais, uma maneira de realizar tal tarefa é utilizando os algoritmos genéticos que auxiliam na captura dos principais recursos que podem otimizar a precisão, tornando a abordagem menos empírica, os recursos escolhidos por esses algoritmos são utilizados como dados de entrada para modelos que são capazes de gerar previsões sendo por fim validado por análises estatísticas para chancelar a efetividade dos modelos preditivos. Os métodos utilizados para a seleção de modelos se dividem em: baseado em modelo e independente de modelo. A diferença entre os métodos é que no caso do método baseado em modelo, o mesmo, abrange a geração de um modelo completo para classificação ou previsão para estimar a performance das variáveis de suporte que são escolhidos por uma métrica, já no caso do método independente de modelo, o mesmo é embasado em soluções estatísticas entre subconjuntos da variável a ser estimada com ajuda dos atributos auxiliares. Por fim o objetivo primordial é elencar uma métrica afim de realizar comparações entre os atributos auxiliares com os atributos a serem estimados pelo modelo, selecionando aqueles de melhores performances sem aplicar o modelo preditivo.

Conforme Bishop (2006), é afirmado que o coeficiente de determinação ($R^2$), e o erro quadrático médio (RMSE) são parâmetros ótimos de avaliação e validação de modelos preditivos.

É descrito por Baykasoglu et al. (2009) que para as barragens que possuem como o modo principal de falha para colapso a liquefação, é de alta complexidade a previsibilidade da ruptura, pois, além de depender de diversos fatores físicos diferenciados, depende também da relação entre eles, a prática tem demonstrado que essas relações são altamente não lineares. Muitas foram as tratativas de se querer realizar previsões de liquefação através de estatística clássica e redes neurais, porém, os resultados não foram satisfatórios, nesse sentido uma nova

abordagem através de classificação é uma diferenciação que pode trazer melhores resultados na precisão e eficácia.

Em seu artigo Xu et al. (2018) comentam que para desastres onde ocorram problemas de deslizamentos de terra, é importante foco em estudos para determinar de forma precoce esse tipo de acidente para salvaguardar vidas e prevenir danos materiais, nesse sentido, é salientado que os modelos que existem para este tipo de problema são estáticos, porém, este tipo de evento é dinâmico, sendo uma boa abordagem consiste em dividir em componentes de tendência e periódicos usando decomposição de modo empírico, onde o componente de tendência é previsto utilizando uma curva S e o componente periódico é previsto utilizando redes neurais de *Long Short-Term Memory* (LSTM), onde, este tipo de modelagem encontrou resultados melhores que os modelos estáticos como vetor suporte a regressão e rede neural de retro propagação.

Segundo Zhang et al. (2015) a precipitação pluviométrica é uma das causas que podem desencadear falhas em estruturas geotécnicas em toda parte do mundo, é afirmado que muitas análises e estudos já foram endereçados no sentido de investigar o efeito de percolação de águas de chuva em estruturas geotécnicas, é importante sempre realizar algumas ações como o estudo de modelos conceituais, análise analítica e modelagem numérica.

De acordo com Khalifah et al. (2020) recentemente as técnicas de ML tem ajudado a disciplina de geociências para endereçar problemas tradicionais de difícil modelagem, foi utilizado nesse trabalho os algoritmos genéticos e redes neurais para prever permeabilidade em estruturas geotécnicas, e destacou que o ML previu com melhor acurácia se comparado com modelos convencionais baseados em estatística clássica, pois, esses tem origem empírica fixa. Os modelos convencionais fazem boas previsibilidades caso haja a utilização de parâmetros de entrada com alta qualidade, porém, os métodos de ML também tem suas limitações e devem ser considerados para que seu uso não seja indiscriminado. Lembrando que as técnicas convencionais e de ML precisam passar pela etapa de treinamento com dados consistentes que representam o problema a ser resolvido. Por fim a validação dos resultados trará qual método é o melhor para cada tipo de solução a ser entregue por uma pesquisa.

Para Mohamed (2016), a engenharia geotécnica lida com estruturas geotécnicas que pelas próprias características intrínsecas apresentam comportamentos diversos com nível de incerteza alto, é salientado que as formações dos materiais não são homogêneas, sendo que muitos modelos matemáticos têm alta taxa de falha para problemas de engenharia geotécnica,

pois, geralmente há grande simplificação do problema ou a incorporação de diversas hipóteses nos modelos. As técnicas de ML para esse tipo de disciplina é mais indicada pelos treinamentos aplicados nos modelos tirando assim a simplificação dos problemas e incorporação de hipóteses, além disso, os modelos de ML podem ser sempre atualizados para obtenção de melhores resultados além do que essas técnicas lidam muito bem com relacionamentos não lineares.

Segundo afirma Puri et al. (2018), na disciplina de engenharia geotécnica são largamente utilizadas correlações empíricas para avaliar estruturas geotécnicas, através de dados de campo ou de laboratório, utilizam se muitos métodos estatísticos para encontrar as diversas correlações possíveis, utilizando o ML que tem capacidade de aprender com dados de entrada e saída, pode se capturar de forma eficiente a relação funcional dos dados históricos de estruturas geotécnicas, mesmo se as relações fundamentais não são conhecidas ou o significado físico é de difícil compreensão.

Através do estudo de trabalhos correlacionados a esse, fica claro que a utilização de métodos de ML tem grandes vantagens em relação a métodos estatísticos convencionais, trazendo maior eficiência e precisão na previsão dos dados futuros de um problema, para tanto serão utilizadas 6 técnicas de ML sendo elas a *random forest, AdaBoost*, Redes Neurais, Regressão Linear, Árvore de Decisão e Stochastic Gradient Descent para regressão e classificação dos dados deste estudo.

Para os estudos de regressão serão avaliados os seguintes parâmetros de desempenho o MSE, RMSE, MAE e $R^2$, sendo que o RMSE será determinante para avaliação do melhor parâmetro de eficiência dos modelos utilizados. A partir do melhor RMSE serão realizados as parametrizações da melhor técnica de ML para o cascateamento entre a regressão e classificação.

De acordo com Almalaq (2017) o MSE (*Mean Squared Error*) é a diferença entre os valores originais e previstos extraídos pelo quadrado da diferença média sobre o conjunto de dados, o RMSE (*Root Mean Squared Error*) é a taxa de erro pela raiz quadrada do MSE. O MAE (Mean Absolute Error) é a diferença entre os valores originais e previstos extraídos pela média da diferença absoluta sobre o conjunto de dados. O $R^2$ representa o coeficiente de quão bem os valores se ajustam em comparação com os valores originais.

Para a classificação os parâmetros de desempenho a serem avaliados serão *Area Under the ROC (Receiver Operating Characteristic) curve* (AUC), a Acurácia, *classification accuracy* (CA), o *F1-Sco*re, a Precisão (*Precision*) e Sensitividade (*Recall*).

Segundo Almalaq (2017), o AUC é o valor que agrega os limiares das taxas de verdadeiro positivo e taxas de falso positivo, a acurácia é a porcentagem de amostras positivas e negativas classificadas de forma correta sobre a soma de amostras positivas e negativas. O *F1-score* é a média harmônica entre a Precisão e o *Recall*. A Precisão é a porcentagem de amostras positivas classificadas corretamente sobre o total de amostras classificadas como positivas. A Sensitividade (*Recall*) é a porcentagem de amostras positivas classificadas corretamente sobre o total de amostras positivas.

# 3. Materiais e métodos

Este capítulo apresenta a etapa de coleta e tratamento dos dados da instrumentação geotécnica que servirão de entrada para os métodos de ML que serão utilizados nessa pesquisa, passando os mesmos pelos processos de treino e validação.

## 3.1. Coleta de dados

O GEOTEC é a plataforma oficial da mineradora Vale S/A para armazenamento de dados históricos de instrumentação geotécnica. Este trabalho considera dados coletados do GEOTEC referentes a valores diários de dois indicadores de nível de água, da barragem Capitão do Mato. Essa estrutura tem as seguintes coordenadas de localização, latitude 7.773.780 e longitude 612.288, a coleta de dados se deu entre os dias 01/01/2020 até 28/05/2020, os indicadores de nível de água que fazem parte desse estudo, são: CMTBCMTNA001 e CMTBCMTNA007, doravante INA001 e INA007 respectivamente, foram totalizadas 149 instâncias para cada indicador de nível de água.

Foram coletadas 149 instâncias para a precipitação pluviométrica de uma estação meteorológica de posição de latitude 7.775.983 e longitude 611.532, totalizando assim para os três instrumentos, 447 instâncias de dados para o *data set* a ser aplicado nas técnicas de ML.

A Figura 7 apresenta a interface do sistema GEOTEC de onde foram extraídos os dados para essa pesquisa.

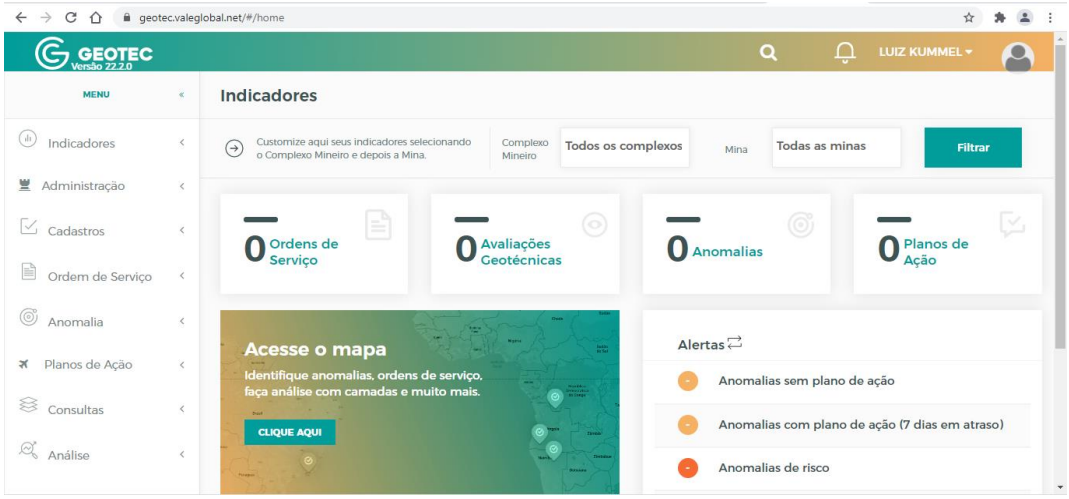


Figura 7: Sistema GEOTEC. Fonte: Vale (2021).

A Figura 8 apresenta a barragem Capitão do Mato com suas seções e a locação dos instrumentos geotécnicos na mesma, a estrutura é dividida em 3 seções verticais e 2 horizontais.


Figura 8: Planta de locação de instrumentos. Fonte: Vale (2018).

O INA001 tem sua localização de instalação na latitude 612.256.87 e longitude 7.773.858.78, sendo que sua cota de topo é 1.173,83, a cota de fundo é 1.165,53, a profundidade de instalação do instrumento é de 8,30 metros.

O INA007 tem sua localização de instalação na latitude 612.187.96 e longitude 7.773.850.31, sendo que sua cota de topo é 1.174,18, a cota de fundo é 1.155,48, a profundidade de instalação do instrumento é de 18,70 metros.

A distância entre a localização da estação meteorológica que mede a concentração pluviométrica e a barragem Capitão do Mato é de 2,32 km, conforme é apresentado na Figura 9.


Figura 9: Distância entre estação meteorológica e barragem de Capitão do Mato. Adaptado. Fonte: Google Earth (2021).

Os indicadores de nível de água escolhidos para o estudo tem comportamentos de leitura distintos em função do furo de instalação que estão localizados, tais comportamentos serão percebidos quando da apresentação gráfica dos mesmos.

O INA001 apresenta leituras com muitos picos de variação, ao passo que, o INA007 apresenta leituras mais suaves, sem o comportamento de leituras bruscas.

Na Figura 10 é apresentada a plotagem do gráfico dos dados obtidos pelo INA001 no período de análise da amostra.

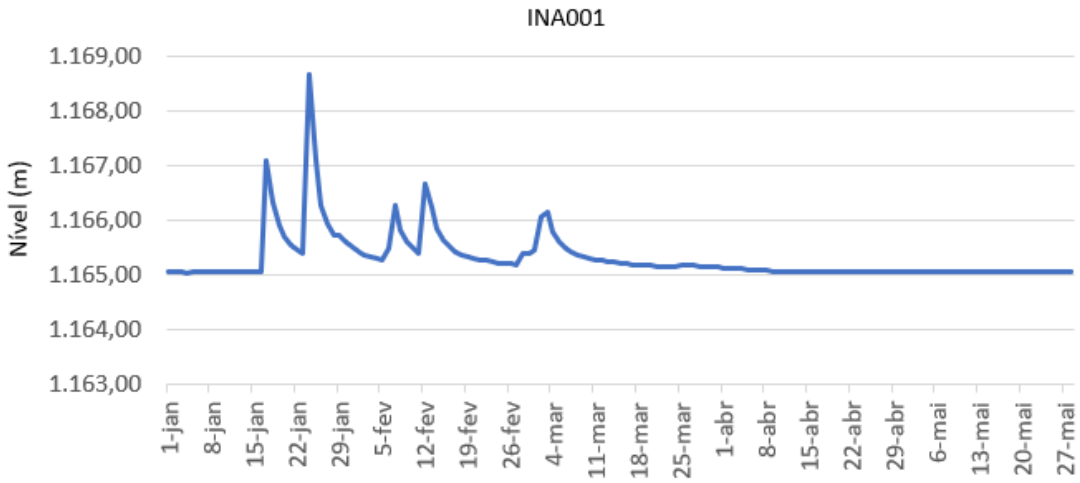


Figura 10: Gráfico de linha do INA001 (01/01/20 até 28/05/20). Fonte: Autor (2021).

O INA001 tem suas leituras do nível do lençol freático muito próximas a cota de fundo de instalação do instrumento.

Na Figura 11 é apresentada a plotagem do gráfico dos dados obtidos pelo INA007 no período de análise da amostra.

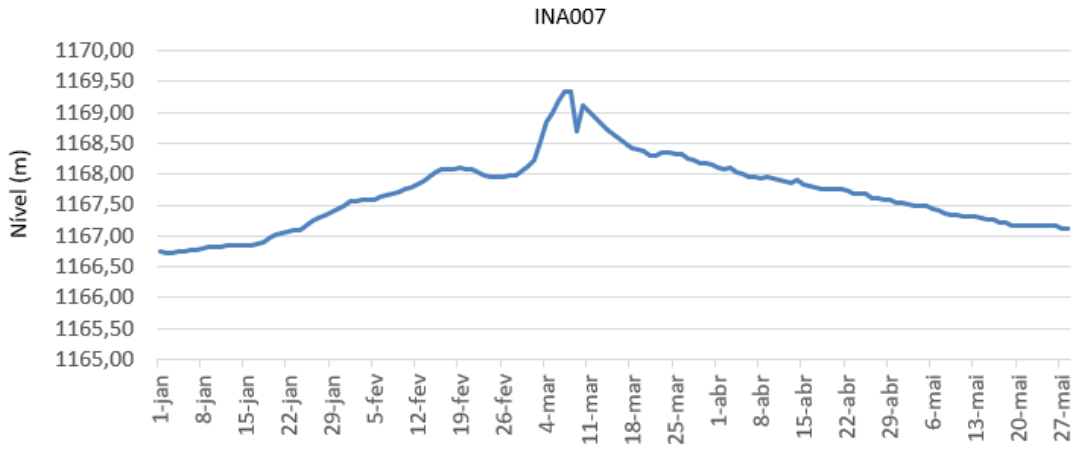


Figura 11: Gráfico de linha do INA007 (01/01/20 até 28/05/20). Fonte: Autor (2021).

O INA007 apresenta leituras do nível do lençol freático longe das cotas de tubo e fundo da instalação do instrumento.

Para efeitos de comparação das leituras dos indicadores de nível de água 001 e 007, foi elaborado um gráfico apresentado na Figura 12, onde são apresentadas as curvas de cada instrumento. Percebe-se que as leituras do INA007 tem patamares maiores que as leituras do INA001, isso ocorre devido à posição de instalação física de cada instrumento na barragem, pois, a posição de latitude e longitude são diferentes, além disso, a profundidade que cada instrumento está inserido é também diferente conforme descrito no item 3.1 deste trabalho.

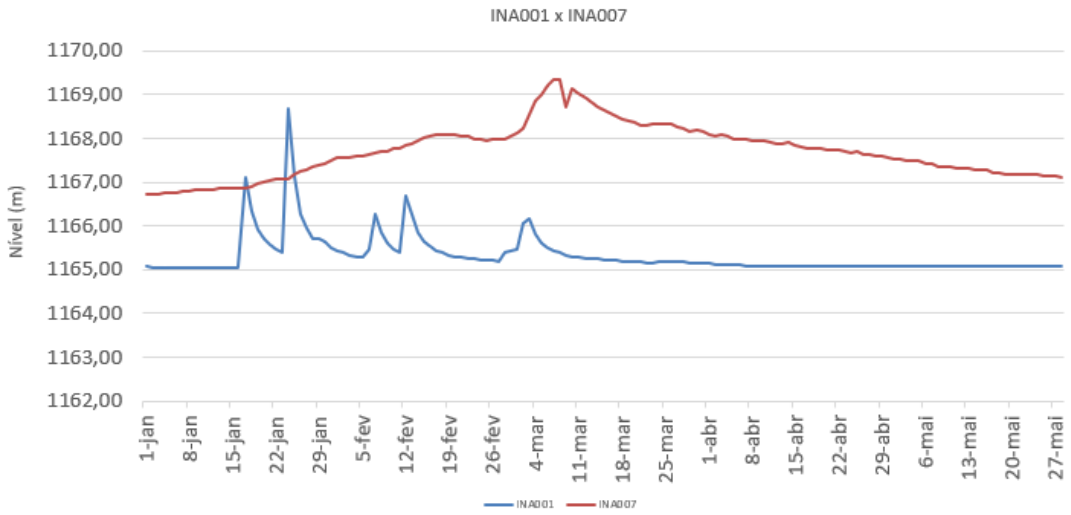


Figura 12: Gráfico dos indicadores de nível de água 001 e 007. Fonte: Autor (2021).

É perceptível na Figura 12 que o comportamento do INA001 é muito mais brusco com mudanças repentinas das medições se comparado ao INA007, isso pode ser explicado pelo fato das leituras do INA001 estarem muito próximas da cota de fundo e longe da cota de topo, o que não ocorre com o INA007 que tem suas leituras longe dessas duas cotas.

A Figura 13 apresenta o gráfico da concentração pluviométrica, em mm.

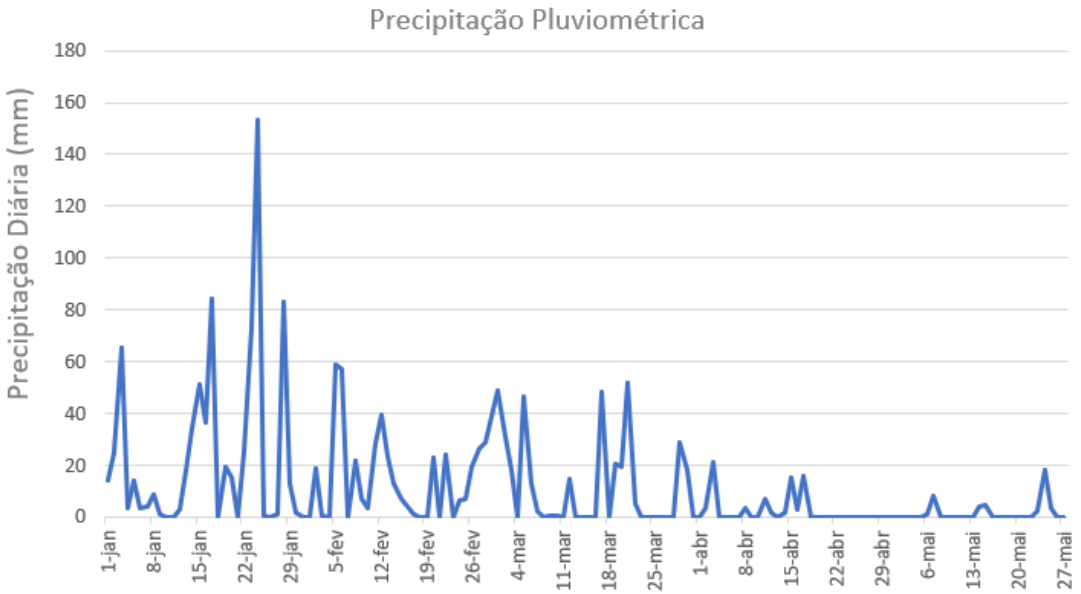


Figura 13: Concentração pluviométrica (01/01/20 até 28/05/20). Fonte: Autor (2021).

A partir da série histórica coletada para análise, foi elaborada a Tabela 1, ela apresenta dados dos valores máximos e mínimos observados bem como as cotas de tubo e fundo de cada indicador de nível de água analisado para o estudo em questão.

Tabela 1: Dados em metros dos indicadores de nível de água 001 e 007. Fonte: Autor (2021).

| | Original | | Transformado (-1160) | |
|---|---|---|---|---|
| | **INA001** | **INA007** | **INA001** | **INA007** |
| Mínimo observado | 1.165,05 | 1.166,71 | 5,05 | 6,71 |
| Máximo observado | 1.168,05 | 1.169,34 | 8,05 | 9,34 |
| Cota do tubo | 1.173,83 | 1.173,57 | 13,83 | 13,57 |
| Cota de fundo | 1.165,53 | 1.154,74 | 5,53 | -5,26 |

Para uma melhor visualização e simplificação foi realizada a subtração do valor de 1160 m, essa operação não interfere em nenhum momento nos resultados que serão obtidos, entretanto, melhora a facilidade de visualização e interpretação dos mesmos.

Houve uma etapa de tratamento de dados, onde foram verificadas se existiam dados faltantes, de fato havia um dado faltante para o INA007 no dia 26/03/20, para tanto foi realizada uma média aritmética simples dos valores vizinhos (25/03/20 e 27/03/20) para que não houvesse dados faltantes para alimentar os modelos de ML. Em cao de dados faltantes os modelos de ML podem ter uma performance insatisfatória.

Foram ajustadas a formatação de separadores de casas decimais para que a entrada do modelo pudesse trabalhar com dados consistentes, além disso, as datas foram colocadas no mesmo formato de DD/MM/AAA. Os dados não foram normalizados entre 0 e 1 para alimentar as entradas dos modelos preditivos de regressão e classificação.

## 3.2. Os Modelos de *Machine Learning*

Diversos modelos de ML podem ser utilizados para regressão e classificação em diversos tipos de problemas que precisam ser solucionados.

Para esse trabalho foram modelados seis métodos de *machine learning* sendo eles:

- *Random Forest*,
- *AdaBoost*,
- Rede Neural,
- Regressão Linear,
- Árvore de Decisão,

- *Stochastic Gradient Descent*.

Todas as técnicas de ML serão descritas nos itens de 3.3 a 3.8 de forma sucinta para entendimento da forma de funcionamento dos mesmos.

### *3.3. Random Forest*

O *Random Fores*t é um método de aprendizado de máquina que utiliza conjuntos de árvores de decisão para realizar regressão ou classificação dependendo do tipo de problema a ser tratado e foi proposto por Breiman (2001), ele foi desenvolvido para solucionar o problema de poda de árvores de decisão, prevenindo dessa maneira o *overfitting*, reduzindo dessa maneira o tempo para construir o modelo multivariado sendo ele um método de aprendizagem supervisionada que trabalha muito bem correlacionando dados não lineares, esse método é capaz de tratar dados faltantes, *outliers* e possui capacidade de tratar classes não balanceadas e variáveis que não exibem uma distribuição normal, conforme afirmam Breiman (2001) e Lee et al. (2013).

Breiman (2001) afirma que essa técnica aplica um conjunto de árvores de decisão sem a poda contando com os métodos de *bagging* e seleção aleatória de variáveis. O *bagging* tem por objetivo treinar cada árvore com um conjunto de dados de treinamento com reposições criadas sem dependência dos dados anteriores. Utiliza se 2/3 das amostras no treinamento para gerar uma árvore, os 1/3 restantes de dados são utilizados para validar a performance o modelo.

Já o processo de seleção aleatória, seleciona de forma aleatória certo número de variáveis para serem utilizadas em cada nó durante a geração da árvore conforme propõem Svetnik et al. (2003), este número de variáveis tem o valor de $\sqrt{(n)}$, para problemas de classificação e 1/3 x $(n)$ para problemas de regressão. No fim o modelo final demonstra boa generalização e robustez em detrimento a amostras com ruídos e anômalas, pois, a utilização desses dois processos aleatórios e a geração das árvores de forma independente corroboram para tal.

Conforme Bauer e Kohavi (1999), a quantidade de árvores é disposta na forma $\{T_1(\theta_1), T_2(\theta_2) \ldots . T_n(\theta_n)\}$, onde $T_i$ é cada árvore do modelo, e $(\theta_i)$ são amostras com reposição de dimensões de subconjuntos de treinamento vezes q, onde a incógnita q é 2/3 x $(n)$. A Figura 14 traz a representação gráfica de uma *random forest*.

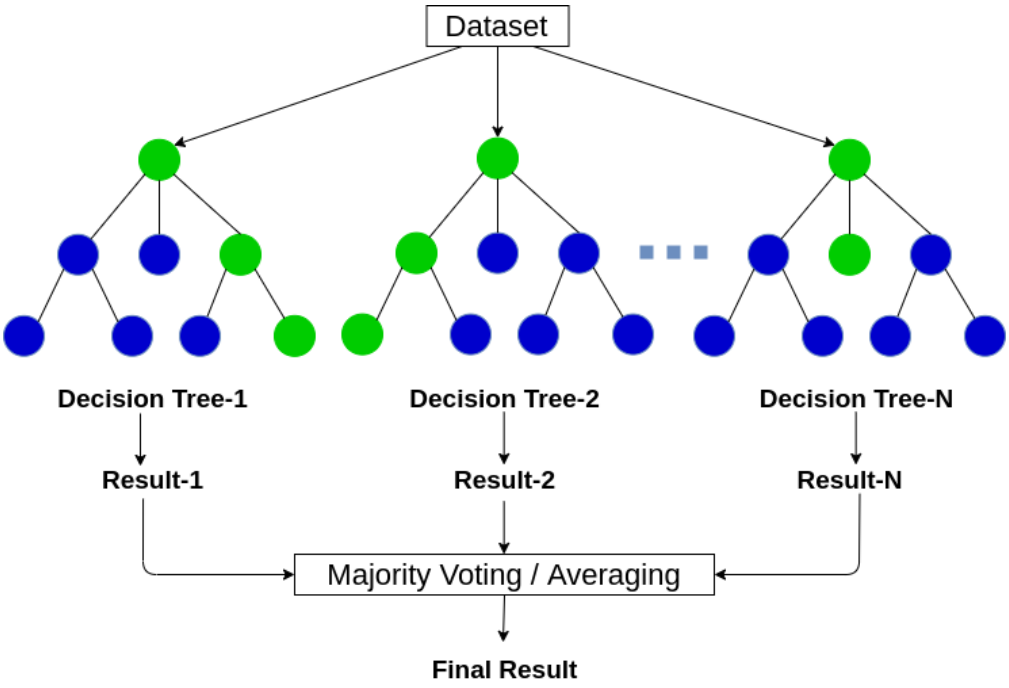


Figura 14: Representação de uma random forest. Fonte: Analytics Vidhya (2020).

Cada árvore do modelo gera uma resposta $y_{1,i}$ para cada uma das amostras alimentadas no modelo, $W\ \{T_1\ (W) = (y_{1,1}), T_2\ (W) = (y_{1,2})\ \ldots, T_B\ (W) = (y_{1,B})\}$, a saída pode ser um resultado de classificação ou de regressão, que é a resposta final do modelo.

O *random forest* conta com três parâmetros a serem configurados, sendo eles: o tamanho mínimo de nó, o número de árvores e por fim o número de variáveis selecionadas aleatoriamente. A configuração do *random forest* se dá da seguinte maneira (1) parametriza se o número mínimo de nós, a boa prática utiliza o valor 1 para problemas de classificação e 5 para problemas de regressão, (2) é o momento de parametrizar o número de variáveis aleatórias, ele varia de 1 a n (número total de variáveis) e (3) a última parametrização a se realizar é a quantidade de árvores do modelo, esse número deve ser o tão grande de modo a estabilizar o erro das amostras, caso o número for muito grande ele não compromete de negativamente os resultados, entretanto, consumirá de forma não otimizada os recursos computacionais para processamento de dados e o tempo da análise será mais prolongado conforme recomendam Svetnik et al. (2003) e Prasad et al. (2006).

### *3.4. AdaBoost*

Conforme descrito por Freund e Schapire (1999), o *boosting* é um método para melhorar a performance de qualquer técnica de ML. É uma técnica que é utilizada de forma combinada com outras técnicas, como, as de redes neurais, árvores de decisão entre outras. O *boosting* executa uma combinação de classificadores criados pelo mesmo algoritmo de ML de forma que seu funcionamento é adaptado conforme os erros cometidos pelo classificador anterior.

O *boosting* é a geração de novos classificadores melhores e mais adaptados aos problemas pelo fato de corrigirem e aumentarem a performance dos classificadores criados de maneira isolada.

Freund e Schapire (1997) comentam que o *AdaBoost* (*Adaptive Boosting*) apresenta propriedades intrínsecas que qualificam o mesmo a ser largamente implementado e utilizado comparando se a algoritmos antecessores como, por exemplo, o Support Vector Machine (SVM).

O algoritmo funciona primeiramente na etapa de treinamento, sendo o conjunto de entrada na forma S= {($x_1$, $y_1$) ($x_2$, $y_2$), ..... ($x_m$, $y_m$)}, onde cada $x_i$ representa o vetor de atributos que é um conjunto de dados referente aos parâmetros a serem avaliados, $y_i$ representa grupo de classificação associado ao $x_i$. E m é o número total de amostras do *data set*.

A base de trabalho do algoritmo é sua repetição por diversas iterações, sendo que cada iteração o *AdaBoost* entrega ao algoritmo uma base de distribuição de pesos referentes a cada um dos dados do treinamento, sendo os pesos envolvidos igualmente distribuídos ($D_0$=1/m).

Em cada ciclo da aprendizagem, o *AdaBoost* gera uma hipótese $h_t$ que está diretamente associada aos pesos atuais com o foco de priorizar a correta classificação dos dados que apresentam os maiores pesos correlacionados.

O propósito do algoritmo é criar uma hipótese para minimizar o erro de treinamento $e_t$.

$$et = \sum i: ht(xt) \neq yt\ Dt(i)^k$$

Na sequência, esses mesmos pesos são reavaliados no sentido de incrementar aqueles que são relacionados as informações que foram classificadas de forma incorreta e uma nova iteração

se inicia. A partir do momento que todas as iterações são realizadas, o algoritmo passa a combinar todas as hipóteses intermediárias para gerar uma hipótese final $H_{(x)}$.

O método para obtenção da hipótese final $H_{(x)}$ disponibilizada pelo *AdaBoost* é a combinação de forma ponderada das saídas das iterações.

A Figura 15 apresenta de forma esquemática o funcionamento do algoritmo *AdaBoost*.

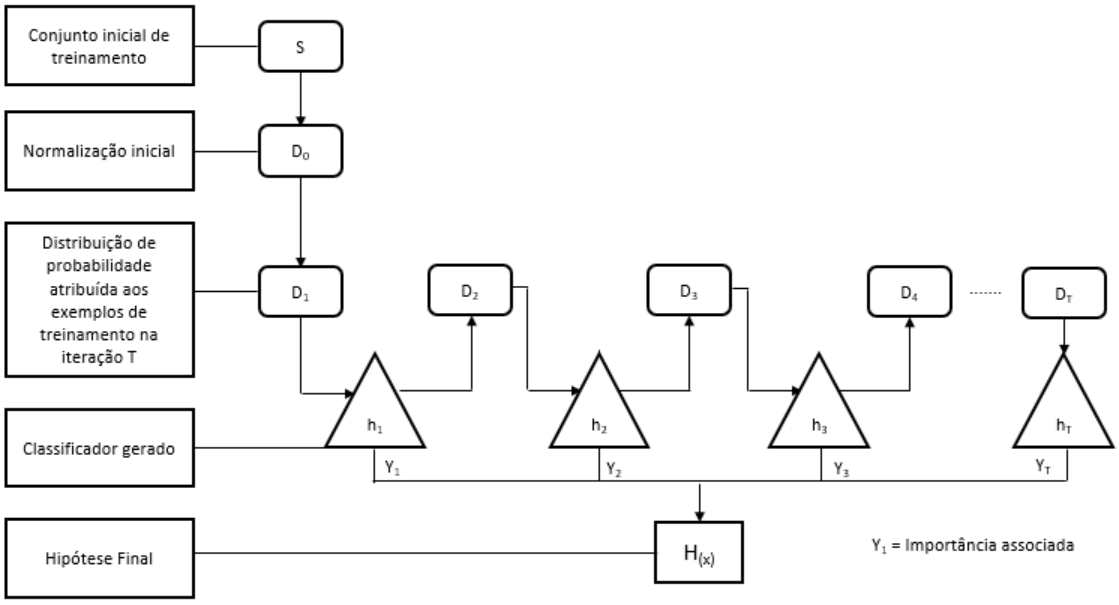


Figura 15: Esquemático de funcionamento do *AdaBoost*. Fonte: Autor (2021).

## 3.5. Redes Neurais

De acordo com Haykin (1999), as redes neurais, chamadas também de redes neurônicas são técnicas de ML que são baseadas no funcionamento do cérebro do ser humano que processa informações complexas, não lineares e paralelas. O cérebro organiza os neurônios em componentes estruturais para execução de processamentos como, por exemplo, o reconhecimento de padrões. As redes neurais são capazes de adquirir conhecimento através de processo de aprendizagem através das forças de conexões entre os neurônios que são chamadas de sinapse, elas que são as responsáveis pelo armazenamento de conhecimento. As redes neurais operacionalizam cálculos sem grandes alterações caso os sinais de entrada forem não lineares.

O modelo das redes neurais é construído por uma grande interconexão de elementos computacionais chamados de neurônios conforme Figura 16. Onde $X_1$, $X_2$, $X_3$, ....., $X_n$ são os dados de entrada do modelo. Os pesos são $W_1$, $W_2$, $W_3$, ....., $W_n$, o b é a junção aditiva

denominada de *Bias* e o σ é a função de ativação da rede neural, a resposta da rede neural é o y.

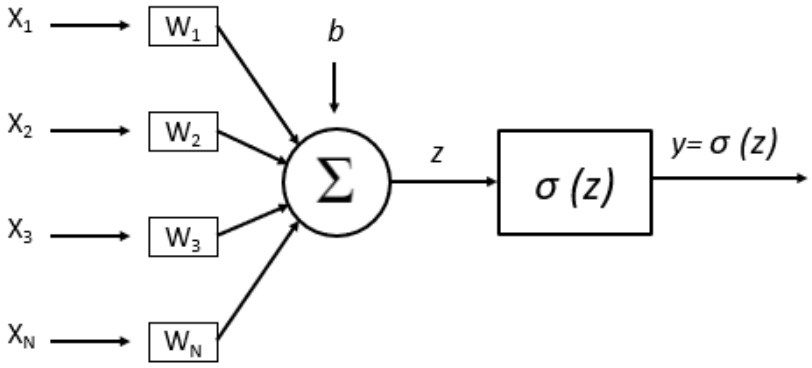


Figura 16: Representação gráfica de redes neurais. Fonte: Autor (2021).

## 3.6. Regressão Linear

De acordo com Maroco (2003), a regressão linear é uma técnica utilizada para modelar relações entre variáveis para predizer o valor de uma ou mais variáveis dependentes a partir de um *data set* de variáveis independentes.

A regressão linear conta com o cálculo de dois coeficientes, o de correlação e determinação, onde o coeficiente de correlação avalia o grau de associação entre duas variáveis, utilizando o coeficiente de Pearson que é o cálculo do quociente entre a covariância entre duas variáveis, por exemplo, x e y pelo produto dos desvios padrão entre essas mesmas duas variáveis.

Se a relação entre as duas variáveis for de $+0,8 \leq R_{xy} < 1$, a mesma é considerada forte positiva, caso a relação seja $+0,5 \leq R_{xy} < +0,8$ é considerada moderada positiva, na condição em que a relação seja $+0,1 \leq R_{xy} < +0,5$ a mesma é considerada como fraca positiva, no caso que a relação seja igual a 0, não há correlação. Da mesma maneira existe a correlação negativa, onde, uma correlação $-0,5 \leq R_{xy} < -0,1$ é considerada fraca negativa, a correlação $-0,8 \leq R_{xy} < -0,5$ é considerada moderada negativa e por fim $-1 \leq R_{xy} < -0,8$ é forte negativa.

O coeficiente de determinação é sugerido para medição e a explicação da reta de regressão, quanto mais próximo de 1 estiver o valor do coeficiente de determinação, maior a porcentagem da variação de Y explicada pela reta predita.

## 3.7. Árvore de Decisão

Segundo Alaboz et al. (2021), a árvore de decisão é um método de ML que utiliza uma função recursiva, ela elabora uma estrutura de árvore de maneira a tornar mais simples e auxiliar na regressão e classificação de amostras não conhecidas. A árvore de decisão subdivide um problema complexo e grande em subproblemas para se tornarem mais simples e mais fáceis de resolução, de forma sequencial cada subproblema também é subdividido para que os mesmos se tornem mais simples.

A Figura 17 apresenta de forma gráfica a árvore de decisão, ela tem o nó raiz que é onde entram os dados do problema a ser avaliado, abaixo do nó raiz, existem os nós de decisões, eles têm por objetivo dividir o atributo avaliado e gerar ramificações, finalmente existem os nós folhas, eles trazem os resultados da árvore de decisão. Cada nó da árvore de decisão é responsável por representar um teste de atributo, desde o nó raiz, passando pelos nós de decisões chegando até o nó folha.

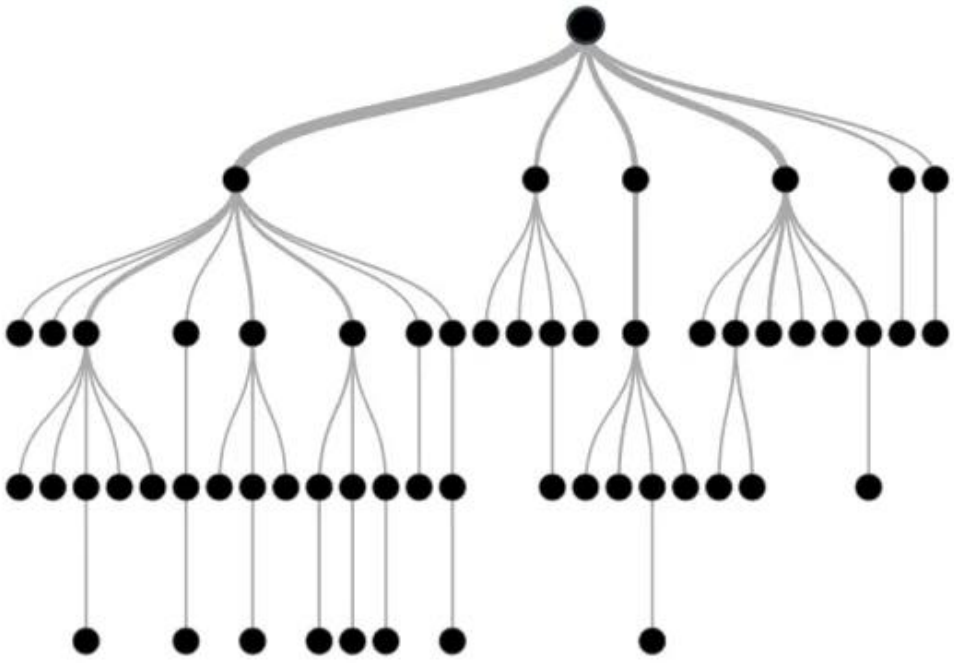

Figura 17: Representação de uma árvore de decisão. Fonte: Vooo (2021).

## *3.8. Stochastic Gradient Descent*

O Gradiente Estocástico Descendente (SGD) é uma técnica de ML onde o seu algoritmo de otimização produz o ajuste de parâmetros de maneira iterativa que tem o propósito de encontrar os valores do coeficiente linear e angular de uma reta que minimizam a função de interesse.

Conforme Maliar e Serguei (2005), a técnica começa preenchendo os coeficientes linear e angular com valores aleatórios, ele otimiza gradualmente a cada iteração, realizando esses preenchimentos de forma mínima incremental até que o algoritmo consiga convergir para um mínimo. Os tamanhos dos incrementos são definidos pelo hiper parâmetro. Caso a taxa de aprendizado seja muito pequena a técnica leva um tempo elevado para convergir devido ao alto número de iterações, porém, se a taxa de aprendizado for alta, a técnica pode ultrapassar o mínimo e dessa maneira não encontrará uma boa solução.

## 3.9. Implementação das técnicas de machine learning

A implementação das técnicas de ML seguem uma sequência para tratamento, pré-processamento, validação e obtenção dos resultados, para tanto é descrito como cada etapa dessa funciona.

A Figura 18 apresenta um fluxo que mostra que o *data set* utilizado, bem como, o *target* são alimentados em um pré-processamento, ele realiza a etapa do tratamento de dados, validando a falta dos mesmos, separadores de casas decimais entre outros, de forma que dados consistentes sejam alimentados na ferramenta de ML.

Na saída do pré-processamento são obtidos os dados de treinamento, o *target* de treinamento e os dados de teste e teste de *target.*

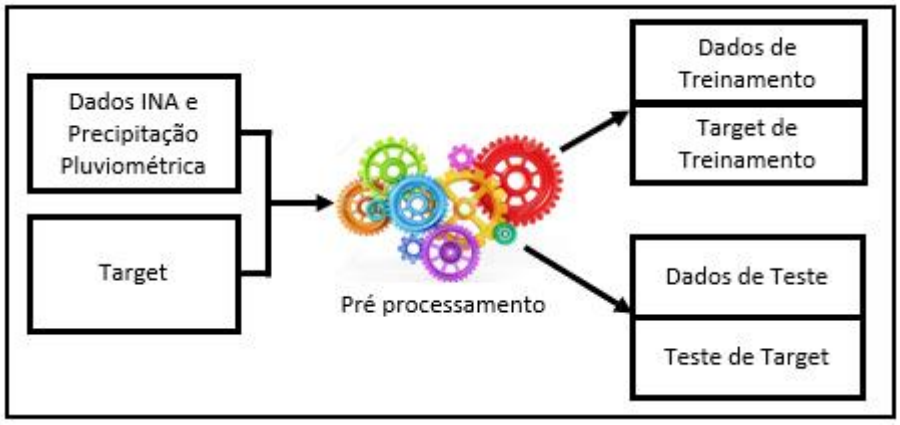


Figura 18: Fluxograma de pré-processamento de dados. Fonte: Autor (2021).

Realizado o pré-processamento, a etapa subsequente é alimentar os dados de treinamento e *target* de treinamento para obtenção da validação cruzada para a avaliação do modelo, como é mostrado na Figura19. Nessa etapa o *data set* é particionado em partes iguais, onde se treina 66% dos dados nos modelos de ML e testa 33%. Esses treinamentos e testes são realizados recorrentemente até que todos os dados sejam processados.

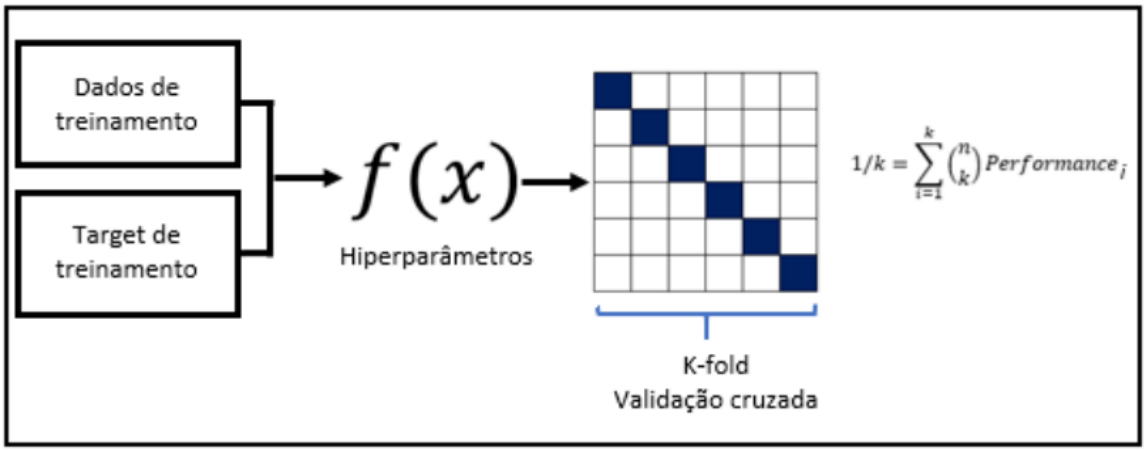


Figura 19: Validação cruzada. Fonte: Autor (2021).

A etapa subsequente é realizar as predições para regressão e classificação onde os valores de predição são calculados e apresentados para elaboração de gráfico do real versus predito. Para a regressão a métrica utilizada será o RMSE e para a classificação será a acurácia, conforme Figura 20.

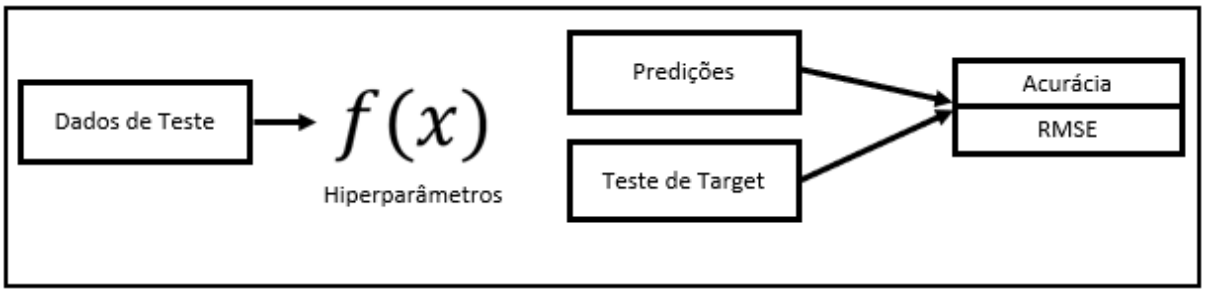


Figura 20: Fluxo de avaliação de modelo ML. Fonte: Autor (2021).

A Figura 21 apresenta a representação do modelo a ser utilizado para a regressão.

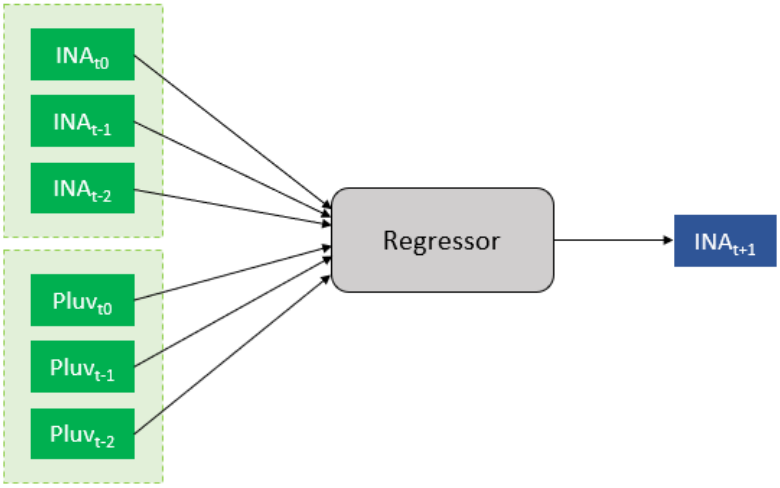


Figura 21: Representação do modelo regressor. Fonte: Autor (2021).

Na entrada do modelo de regressão são inseridos os dados das medições das grandezas físicas do indicador de nível de água e pluviômetro, são seis entradas distintas que utilizam janela de dados de 2, 5 e 10 dias desses instrumentos.

Para a classificação a Figura 21 apresenta a representação do modelo a ser utilizado.

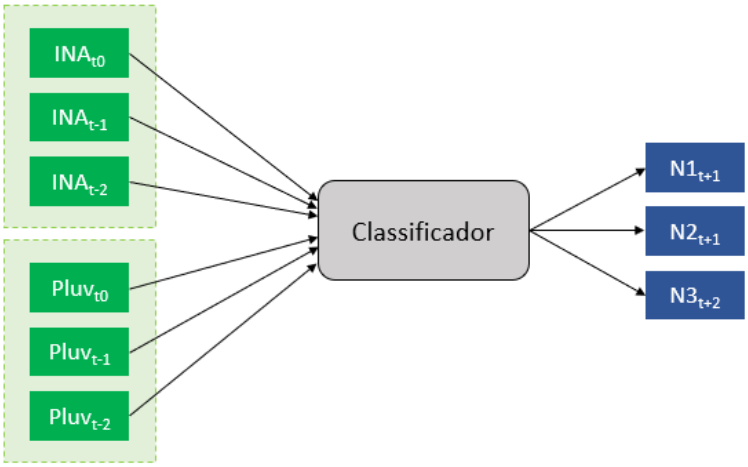


Figura 22: Representação do modelo classificador. Fonte: Autor (2021).

No cascateamento de modelo de regressão será utilizado o modelo de classificação da Figura 22, na entrada desse modelo serão utilizados os dados do indicador de nível de água e dados do pluviômetro.

## 3.10. Modelagem em *Orange Canvas*

Para esse trabalho, a implementação de todas as técnicas de ML foi realizada através do *software* de programação computacional para mineração de dados e aprendizado de máquina denominado *Orange Canvas version 3.30.2*, conforme Demšar et al. (2004). Esse software abrange ferramentas para preparação de dados, classificação, regressão, clustering, mineração de regras para associação e visualização, além de oferecer *scripts* para geração de protótipos de programação visual.

A Figura 23 apresenta um *screen shot* da plataforma *Orange Canvas* onde foram modeladas as técnicas de ML para o problema deste trabalho. Após a validação das informações do *data set* os dados foram alimentados nas entradas dos modelos de ML. Os dados de entrada são dos indicadores de nível de água 001 e 007 e dados de pluviometria. Estes dados além de serem as entradas das técnicas de ML, são levados ao bloco de *test and score* que é responsável por realizar as etapas de treino e teste dos dados, trazendo os resultados de performances de cada técnica de ML que será avaliada através da Tabela 3.

A saída do bloco *test and score* alimenta o bloco denominado predições, ele apresenta os resultados das predições, a partir dele, serão avaliados e gerados os gráficos dos resultados reais versus preditos.

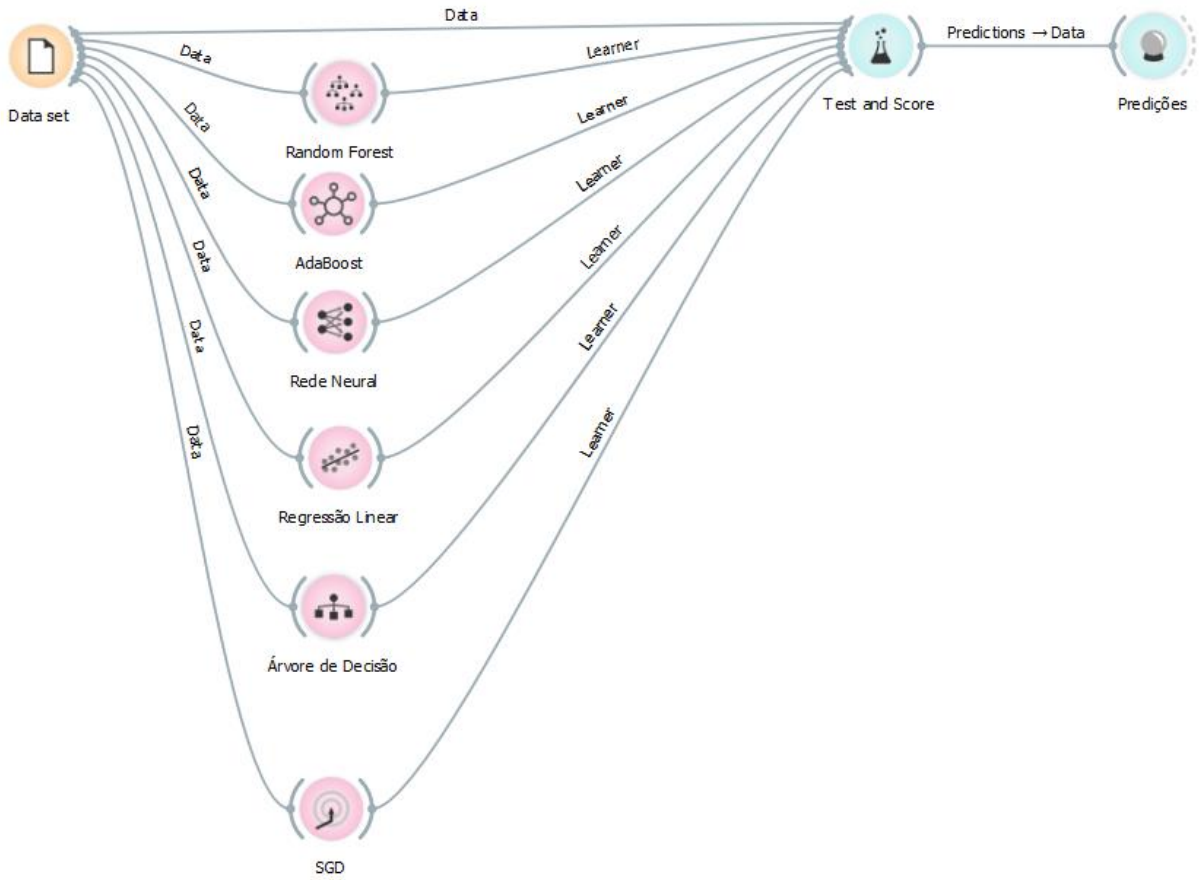


Figura 23: *Screen shot Orange Canvas*. Fonte: Autor (2021).

# 4. Resultados e Discussão

Essa seção tem como objetivo discutir os resultados encontrados no estudo proposto em relação aos objetivos do trabalho.

Este trabalho visou aferir a aplicabilidade dos métodos de ML em cascata para regressão e classificação no estabelecimento de valores de controle para os dados da instrumentação de barragens de rejeito, através da análise das séries históricas das medições dos indicadores de nível de água e pluviômetro.

Será realizada confrontação dos dados históricos reais contra os dados que os modelos de ML fornecerem como resultados de previsibilidade dos indicadores de nível de água, isso é de grande relevância para avaliar se o modelo está com assertividade desejada, no sentido do controle de segurança continuado para estruturas geotécnicas conforme recomendam Menga et al. (1999).

A primeira análise realizada foi a de gráficos de dispersão, calculando se a correlação *Pearson* entre os dados dos indicadores de nível de água 001, 007 e concentração pluviométrica.

Os diagramas de dispersão são gráficos de dados emparelhados (x, y), que podem ser utilizados para analisar padrões e correlações entre variáveis, sendo a interpretação subjetiva, segundo Wang et al. (2016). Eles podem ser utilizados na análise preliminar dos dados da instrumentação, de forma a permitir a seleção dos componentes a serem introduzidos nos modelos estatísticos, são ainda utilizados na determinação de pontos muito afastados dos demais ou atípicos que podem comprometer a análise dos resultados. Os mesmos podem ter as seguintes formas de correlações:

- Positiva: este tipo de correlação acontece quando há uma tendência crescente entre os pontos. Conforme uma variável aumenta, a outra variável também aumenta proporcionalmente.
- Negativa: essa correlação é quando se concentram em uma linha decrescente. Conforme uma variável aumenta, a outra diminui.
- Perfeita: Ela é identificada como perfeita quando não há dispersão entre os pontos, a correlação será total entre os dados, independente da tendência, seja ela positiva ou negativa.

As correlações podem ser categorizadas em fortes e fracas, sendo:

- Forte: Quanto menor for a dispersão dos pontos, maior será a correlação entre eles. Com isso, pode-se identificar como forte quando os dados estão bem próximos e altamente concentrados.
- Fraca: Inversa a correlação forte, então quanto maior for a dispersão dos pontos, menor será o grau de correlação entre os dados, ou seja, eles quase não possuem uma correlação.

A Figura 24 apresenta o gráfico de dispersão de dados entre o INA001 e precipitação pluviométrica.

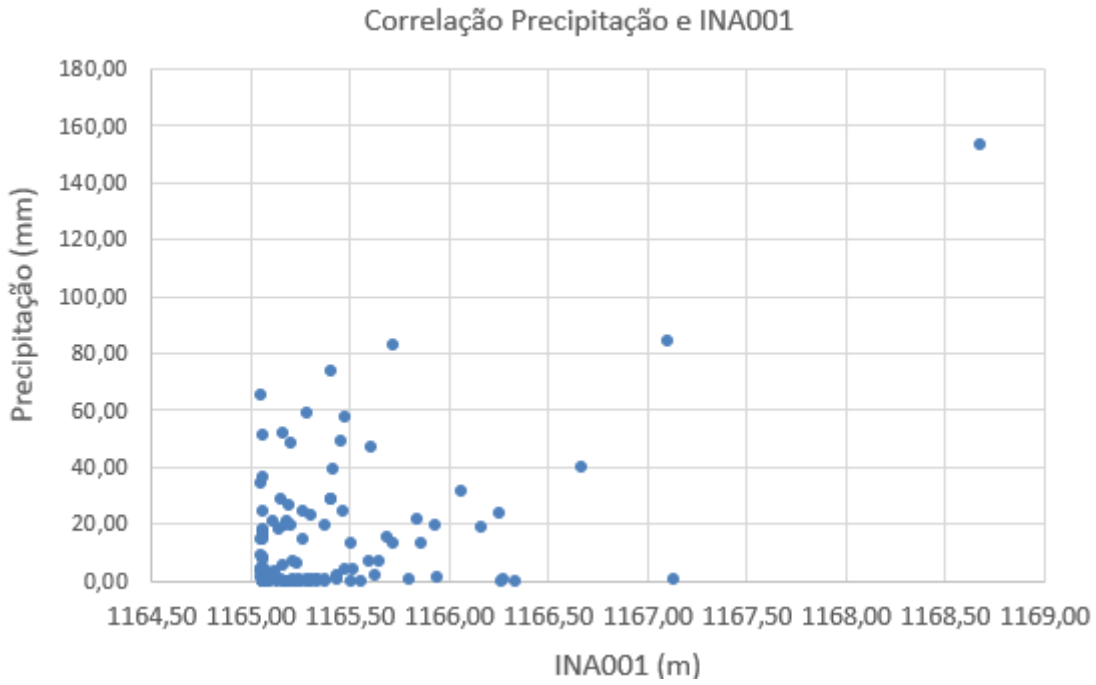


Figura 24: Correlação entre INA001 x Precipitação Pluviométrica. Fonte: Autor (2021).

A Figura 25 apresenta o gráfico de dispersão de dados entre o INA007 e precipitação pluviométrica.

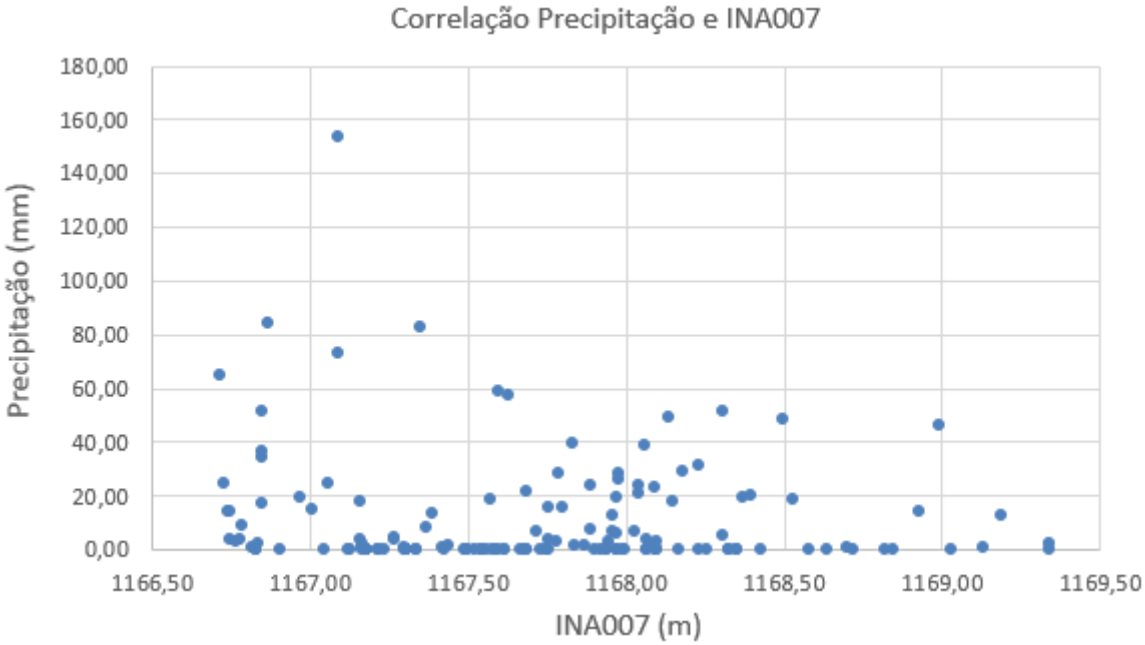


Figura 25: Correlação entre INA007 x Precipitação Pluviométrica. Fonte: Autor (2021).

A Figura 26 apresenta o gráfico de dispersão de dados entre o INA001 e INA007.

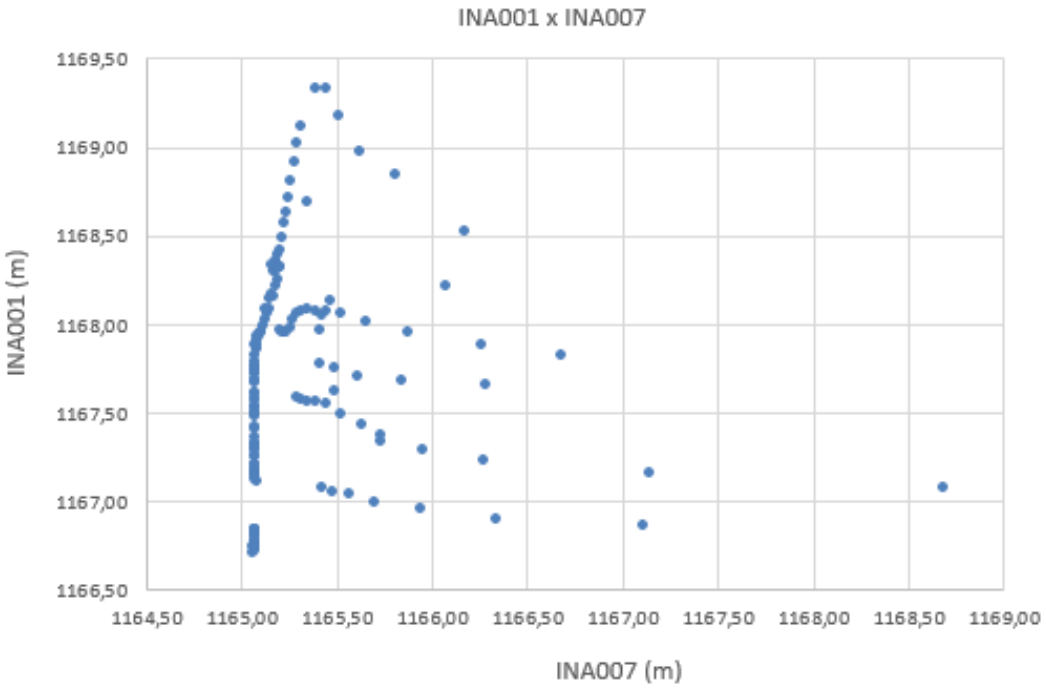


Figura 26: Correlação entre INA001 x INA007. Fonte: Autor (2021).

Conforme Edelmann et al. (2021), a correlação de *Pearson* mede o grau da correlação linear entre duas variáveis quantitativas com dois sinais contínuos que co-variam ao longo do tempo e indicam a relação linear como um número entre -1 (correlacionado negativamente) a 0 (não correlacionado) a 1 (perfeitamente correlacionado).

A Tabela 2 apresenta os resultados da correlação de *Pearson* entre os instrumentos elencados nesse estudo.

Tabela 2: Correlação de Pearson entre INA001, INA007 e Precipitação Pluviométrica. Fonte: Autor (2021).

| | INA001 | INA007 | Precipitação Pluviométrica |
|---|---|---|---|
| INA001 | 1 | 0 | 0,53 |
| INA007 | 0 | 1 | -0,12 |
| Precipitação Pluviométrica | 0,53 | -0,12 | 1 |

Os resultados encontrados nas correlações apontadas na Tabela 2, se demonstraram fracas para o INA001 x INA007 e Pluviometria x INA007.Houve uma moderada correlação entre pluviometria x INA001.

Esses resultados corroboram no sentido de descartar o problema a ser analisado por este estudo pela estatística clássica e realizar as modelagens com ferramentas de ML que tratam bem correlações fracas.

A segunda análise estatística realizada, foi através do gráfico que traz uma visualização da análise exploratória do estudo, é o boxplot, onde é certificada a distribuição de todos os dados e *outliers* da amostra adotada.

A Figura 27 apresenta o gráfico boxplot para os indicadores de nível de água 001 e 007 respectivamente.

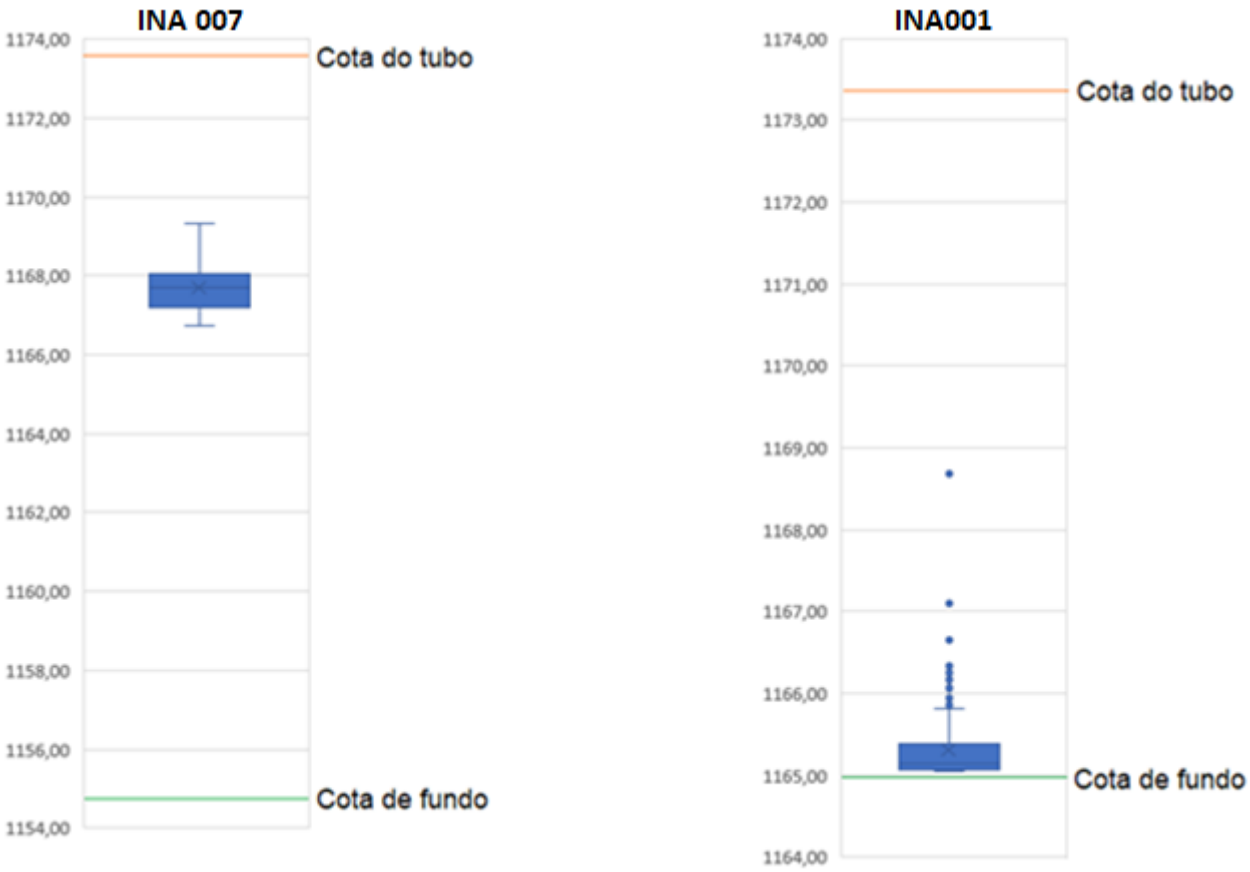


Figura 27: Gráfico boxplot para as amostras dos dados dos indicadores de nível de água 007 e 001.
Fonte: Autor (2021).

Por meio da Figura 27 é percebido que os dados observados no INA007 estão sempre a mais de 2 m dos limites inferiores ou superiores do sensor. Esta é uma observação que ajuda a entender o comportamento mais suave dos dados no sensor, apresentados na Figura 11.

Para o INA001, é percebido que muitos dos dados estão muito próximos ao limite inferior, ou seja, da cota de fundo, o que explica as variações bruscas deste sensor, sendo que 79 dos 149 dados estão dentro de 20 cm de diferença (53,0%), 41 dos 149 dados estão entre 20 e 50 cm (27,5%) e 29 dos 149 dados acima de 50 cm de diferença (19,5%).

## 4.1. Modelos preditivos

Após serem avaliados os gráficos de linha e boxplot, o estudo propõe duas abordagens a serem implementadas. A primeira é uma abordagem de modelo de predição baseado em regressão e a outra é de cascateamento de modelos preditivos de classificação e regressão respectivamente para tratar os objetivos do trabalho.

Na primeira abordagem como exposto será realizada a regressão para estimar valores futuros dos indicadores de nível de água 001 e 007 com os próprios dados desses instrumentos sem inserção da pluviometria, o que caracteriza o problema em um modelo univariado.

Foram avaliados e utilizados seis modelos de ML nesse estudo, conforme descrito no tópico 3.2 deste trabalho.

São apresentados os resultados de MSE, RMSE, MAE e $R^2$ na Tabela 3. Nesse estudo o parâmetro utilizado como de melhor performance para avaliação dos resultados é o erro médio quadrático (RMSE), que quanto menor melhor, é confirmado na Tabela 3 que o modelo que teve o menor RMSE foi o *Random Forest* com o valor de 0,088.

Tabela 3: Resultados dos métodos de ML. Fonte: Autor (2021).

| Método de ML | MSE | RMSE | MAE | $R^2$ |
|---|---|---|---|---|
| RF | 0,008 | **0,088** | 0,051 | 0,978 |
| AdaBoost | 0,008 | 0,091 | 0,048 | 0,977 |
| SGD | 0,013 | 0,112 | 0,056 | 0,965 |
| Árvore de Decisão | 0,013 | 0,115 | 0,069 | 0,963 |
| Regressão Linear | 0,029 | 0,170 | 0,062 | 0,919 |
| Rede Neural | 0,224 | 0,474 | 0,458 | 0,372 |

Diante dos resultados encontrados na Tabela 3 foram realizadas parametrizações diferentes na técnica de RF para verificação e avaliação dos resultados em função das mudanças dos parâmetros.

Para o INA007, foi utilizada a técnica de treino e teste de validação cruzada *k-fold*.

As parametrizações realizadas no *Random Forest* são as janelas de análises de 2, 5 e 10 dias e o número de árvores de decisão aplicadas ao modelo que variam de 20, 50 e 100 árvores.

A Tabela 4 apresenta os resultados em função do número de árvores e janelas de dados.

Tabela 4: Resposta da avaliação do modelo. Fonte: Autor (2021).

| Número de árvores | W2 | W5 | W10 |
|---|---|---|---|
| 20 | 0,105 | 0,100 | **0,088** |
| 50 | 0,104 | 0,100 | 0,089 |
| 100 | 0,106 | 0,099 | 0,090 |

Como pode ser percebido, através dos resultados obtidos pela Tabela 3, o melhor resultado é quando modelo *random forest* tem 20 árvores com a janela de 10 dias.

É observado na Figura 28, os resultados de RMSE das iterações realizadas, confirmando que uma janela de 10 dias e 20 árvores trouxe os melhores resultados diante das janelas de 2 e 5 dias. Com o aumento de árvores esse modelo *random forest* generaliza os resultados, pois, há uma tendência do próprio modelo decorar resultados.

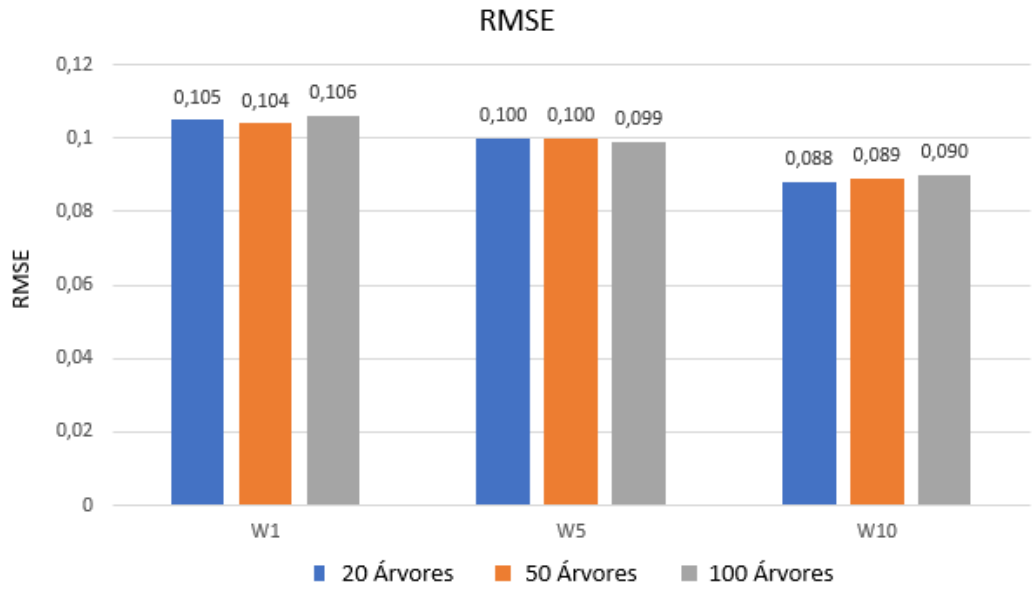


Figura 28: RMSE para predição do valor do INA007. Fonte: Autor (2021).

A partir dos resultados obtidos das parametrizações realizadas no RF, foram coletados os dados de predição para elaborar o gráfico da Figura 29, esse gráfico apresenta as curvas dos dados reais versus os dados preditos.

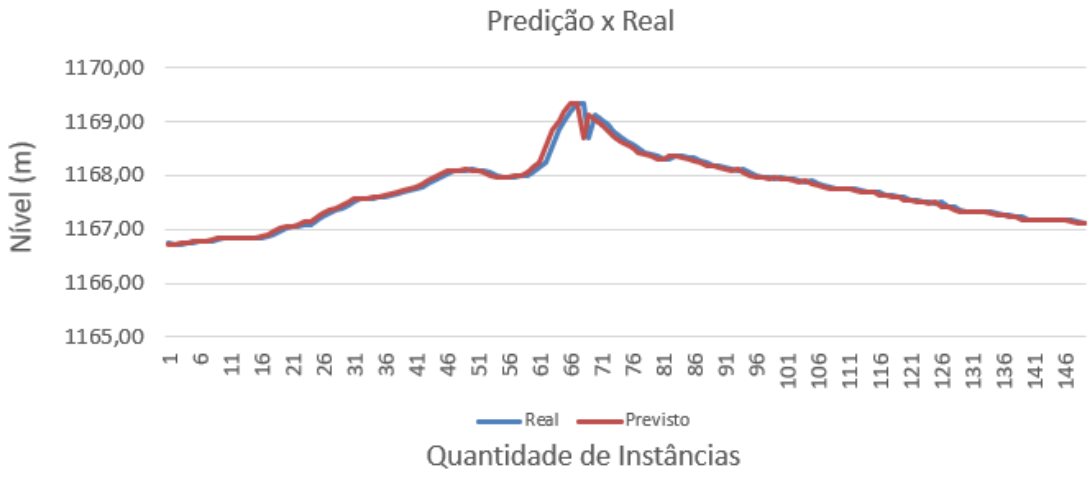


Figura 29: Gráfico valor real x valor estimado para INA007. Fonte: Autor (2021).

Em uma observação visual da Figura 29 é perceptível que os dados preditos têm a mesma forma de onda dos dados reais com pequenos desvios de magnitude.

O erro máximo para este conjunto de dados foi de 48 cm. O erro ficou igual ou acima de 20 cm em 7 das 149 instâncias (4,7%).

Os mesmos procedimentos que foram realizados para o INA007, foram aplicados para o INA001, porém, nesse caso houve a utilização dos dados da precipitação pluviométrica, pois, há uma moderada correlação *Pearson* positiva entre o INA001 e os dados de concentração pluviométrica. Os resultados encontrados para os 6 métodos de MF são apresentados na tabela 5.

Tabela 5: Resultados dos métodos de ML. Fonte: Autor (2021).

| Método de ML | MSE | RMSE | MAE | $R^2$ |
|---|---|---|---|---|
| RF | 0,103 | **0,305** | 0,113 | 0,523 |
| AdaBoost | 0,125 | 0,353 | 0,115 | 0,424 |
| SGD | 0,129 | 0,359 | 0,130 | 0,405 |
| Árvore de Decisão | 0,132 | 0,363 | 0,129 | 0,390 |
| Regressão Linear | 0,165 | 0,406 | 0,144 | 0,239 |
| Rede Neural | 0,226 | 0,475 | 0,156 | -0,044 |

O melhor método de ML para a predição dos dados do INA001 foi também o *Random Forest* que apresentou os melhores resultados na métrica de RMSE tal qual aconteceu com o INA007.

Da mesma forma, foram realizadas as parametrizações da RF com as janelas de 2, 5 e 10 dias e da quantidade de árvores, sendo de 20, 50 e 100. A Tabela 6 apresenta os resultados encontrados com as mudanças realizadas.

Tabela 6: Resultados de iterações para INA001 e pluviometria. Fonte: Autor (2021).

| Número de árvores | W2 | W5 | W10 |
|---|---|---|---|
| 20 | 0,330 | 0,327 | 0,341 |
| 50 | 0,311 | 0,312 | 0,363 |
| 100 | **0,305** | 0,313 | 0,351 |

Para o INA001, o modelo regressor que apresentou o melhor erro médio quadrático (RMSE), foi a parametrização que utilizou a janela de 2 dias de janela de dados históricos e 100 árvores, sendo que resultado obtido foi de 0,305, nesse caso a menor janela de dados e maior

número de árvores apresentou o melhor resultado, porém, aquém do resultado obtido para o INA007.

Dessa maneira o modelo regressor não obteve resultados satisfatórios, pois, o RMSE está alto.

É observado na Figura 30 os resultados de RMSE das iterações realizadas, confirmando que uma janela de 2 dias e 100 árvores trouxe os melhores resultados diante das janelas de 2 e 5 dias. Com o aumento da janela de dias esse modelo *random forest* generaliza os resultados, pois, há uma tendência do próprio modelo decorar resultados.

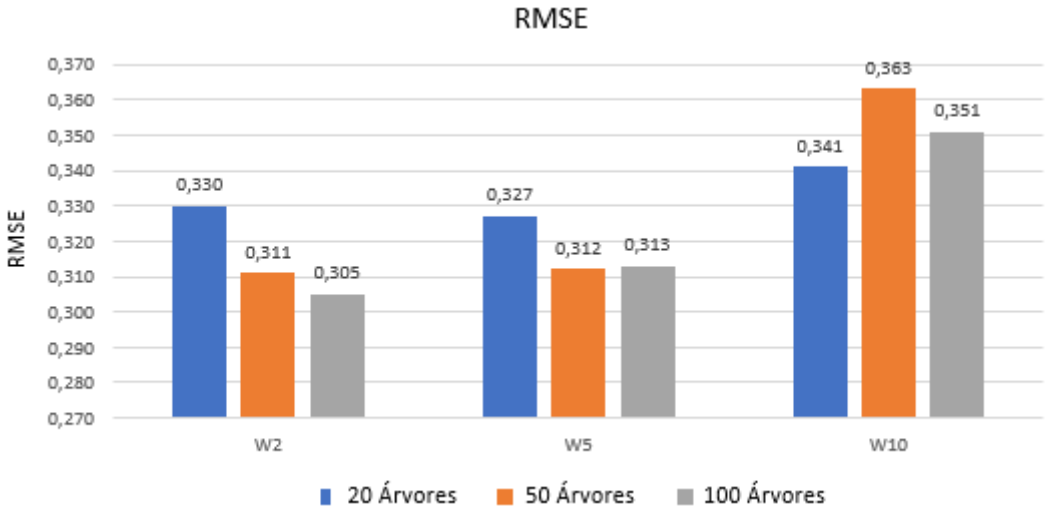


Figura 30: RMSE para predição do valor do INA001. Fonte: Autor (2021).

Foi plotado o gráfico onde são verificados os dados reais versus os dados de previsão calculados pelo modelo preditor, a Figura 31 apresenta o resultado.

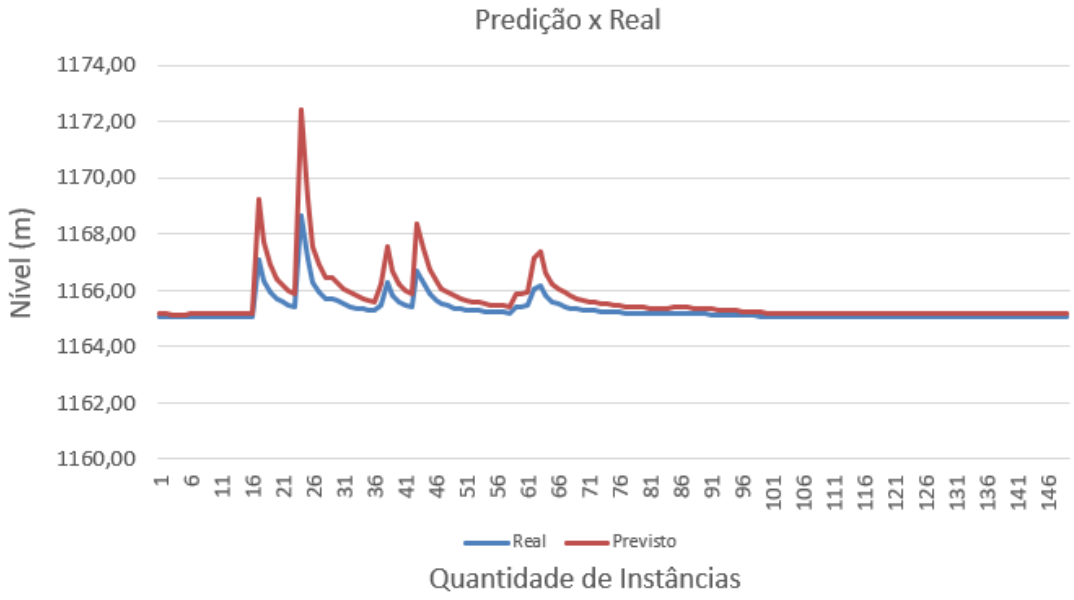


Figura 31: Gráfico valor real x valor estimado. Fonte: Autor (2021).

A Figura 31 apresenta gráfico do valor real versus o valor estimado pelo modelo para o INA001, percebe-se visualmente que a forma de onda do valor predito é muito parecida com a dos valores reais, porém, existem diferenças de magnitudes relevantes observadas. O erro máximo observado para este conjunto de dados foi de 2,51 m, contra 42 cm no caso do INA007. O erro ficou acima de 20 cm em 22 das 149 instâncias, ou seja, erro acima de 20 cm em 14,7% das instâncias. O erro foi superior a 50 cm em 8 das 149 instâncias (5,4% dos casos).

Diante dos resultados insatisfatórios do modelo de predição para o INA001, foi realizada a abordagem de cascateamento de modelos predição, sendo primeiro a classificação e em sequência a regressão.

Essa abordagem, primeiro considera um modelo de classificação pela ferramenta *random forest*, a Figura 22 mostra a representação do modelo do classificador.

O modelo classificador trabalhará como entrada as informações de janelas de 2, 5 e 10 dias das medições geotécnicas do INA001. Como saída o resultado do modelo apresentará 3 classificações para o indicador de nível de água, em função do que o mesmo processar.

As classificações que o modelo apresentará como saída, são 3 classes, apresentadas na Tabela 7.

Tabela 7: Parâmetros de classificação. Fonte: Autor (2021).

| Classificação | Parâmetros (cm) |
|---|---|
| Classe 1 (N1) | Inferior a 20 (19.999) |
| Classe 2 (N2) | Entre 20 e 50 (49.999) |
| Classe 3 (N3) | Acima de 50 |

Foram definidos 3 parâmetros de classificação a serem aplicados no *data set* a ser processado pela ferramenta de ML *random forest*, sendo elas a classe N1 que classifica medições com até 20 cm de variação de erro, a N2 que classifica variações entre 20 até 50 cm de erro e N3 que classifica variações acima de 50 cm de erro.

A Figura 32 apresenta a modelagem realizada no software *Orange Canvas* para coleta dos resultados da matriz de contingência, mais popularmente conhecida como matriz de confusão. Como a modelagem em *Random Forest* foi a que trouxe os melhores resultados na regressão a mesma foi utilizada para o modelo de classificação.

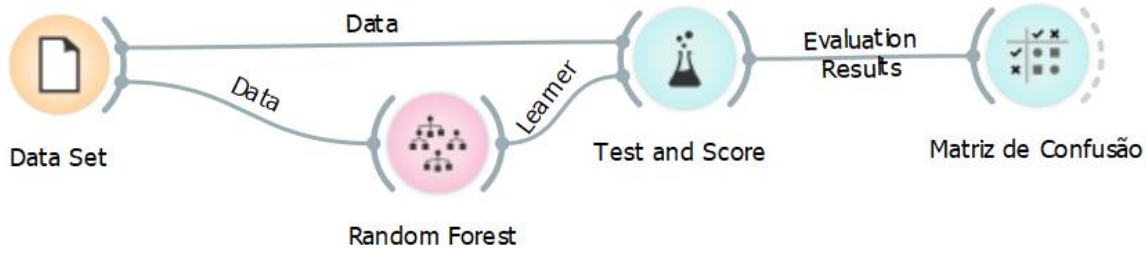


Figura 32: Modelagem da matriz de confusão. Fonte: Autor (2021).

O modelo da Figura 32 é alimentado com o *data set* do INA001, as classes foram definidas e apresentadas na Tabela 7, ligando se o *data set* ao bloco de *Random Forest* e no *test and score* os resultados são elaborados e mostrados através do bloco matriz de confusão.

Os resultados do test and score para a classificação foram: 1.000 para AUC, 0,987 para CA, 0,987 para F1, 0,987 para *Precisio*n e 0,987 para *Recall*. Estes resultados tiveram alta performance o que abona e corrobora a utilização da classificação para o problema.

Os resultados do modelo de classificação baseado em *random forest* são apresentados na Tabela 8.

A modelagem com as séries históricas do INA001 teve o resultado de nível de acerto de 98,7%, ou seja, das 149 instâncias alimentadas no modelo, o mesmo acertou 147. A matriz de confusão na Tabela 8, apresenta o resultado obtido.

Tabela 8: Resultado da matriz de confusão para INA001. Fonte: Autor (2021).

| | N1 | N2 | N3 |
|---|---|---|---|
| N1 | 79 | 0 | 0 |
| N2 | 0 | 28 | 1 |
| N3 | 1 | 0 | 40 |

A matriz de confusão apresenta resultados que abonam o uso da classificação via *random forest* para o estudo elencado.

Após realizar o modelo de classificação é realizado cascateamento para o modelo preditivo de regressão. Novamente é utilizada a ferramenta de ML *random forest* para a previsão de valores do INA001.

A Figura 21 apresenta representação do modelo preditivo com seu regressor. Nesse modelo são inseridos dados de pluviometria para observar a resposta do modelo. Salientando que são seis entradas distintas que utilizam dados históricos de dados de 2, 5 e 10 dias dos instrumentos.

Foram realizadas iterações com os dados advindos do modelo de classificação para prever os valores do INA001, os resultados obtidos podem ser verificados na Tabela 9.

Tabela 9: Resultados de iterações. Fonte: Autor (2021).

| Número de Árvores | Usando INA [w10] e acumulado precipitação pluviométrica de 3, 5 e 8 dias (MSE) | Usando histórico INA001 w10 (MSE) | Usando acumulado de 8 dias (INA001 +Pluv) (MSE) |
|---|---|---|---|
| 20 | 0,321 | 0,333 | 0,333 |
| 50 | **0,309** | 0,328 | 0,332 |
| 100 | 0,313 | 0,331 | 0,330 |

O *random forest* para a parte de regressão foi alimentado com três conjuntos de dados distintos, sendo eles: (1) Dados do INA001 com janela de dados de dez dias e dados de pluviometria acumulados de 3, 5 e 8 dias. (2) Dados somente do INA001 com janela de 10 dias. (3) Dados do INA001 e concentração pluviométrica com janela de 8 dias.

O melhor resultado obtido para o RMSE foi na segunda iteração, utilizando os dados do INA001 com janela de dados de dez dias e dados de pluviometria acumulados de 3, 5 e 8 dias, o resultado foi de 0,309.

Essa análise multivariada de dados pelo ML, trouxe uma boa precisão de resultados preditos onde o erro de classificação para as 149 instâncias alimentadas no modelo ficou entre 0 e 2 nas iterações realizadas, sendo a média de 1,16 erros para as iterações de treino e teste.

A análise multivariada levou em conta previsão após classificação de dados que apresentaram erros acima de 50 cm do limite, utilizando o modelo de previsão chegou se ao seguinte resultado de RMSE de 0,2223.

A Figura 33 apresenta o gráfico do valor predito versus o valor real para os erros acima de 50 cm.

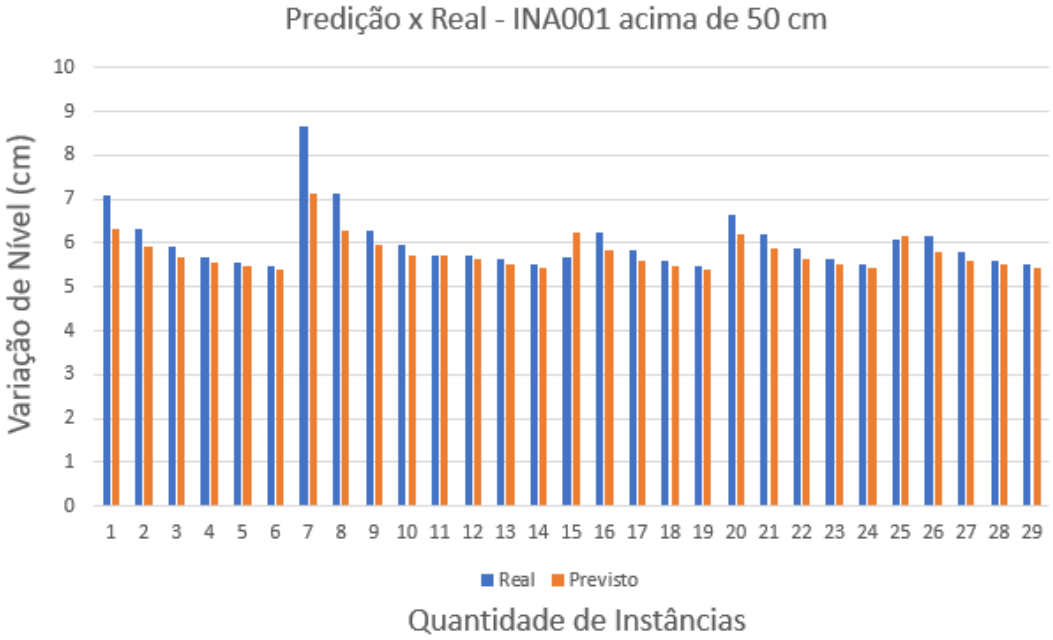


Figura 33: Gráfico predito x real acima com erro acima de 50 cm. Fonte: Autor (2021).

Como pode ser constatado pelo gráfico da Figura 33, o modelo apresentou apenas um erro (em 29 instâncias, [3,4%]) com erro de 0,96 m, para os demais o mesmo ficou abaixo de 20 cm.

# 5. Conclusão

Esse capítulo tem por objetivo realizar os comentários de conclusão deste trabalho, serão respondidas as questões de pesquisa e os objetivos gerais e específicos da pesquisa.

A primeira questão de pesquisa foi: Qual a capacidade de predição, usando apenas histórico para indicadores de nível de água com comportamentos suaves e abruptos?

Em resposta ao primeiro questionamento pode-se afirmar que, a capacidade de predição utilizando apenas os valores dos próprios indicadores de nível de água é muito satisfatório para o instrumento INA007, que tem suas medições longe das cotas de tubo e fundo. O RMSE calculado para esse instrumento foi de 0,088.

Para o INA001 foi observado que suas medições são muito próximas a cota de fundo, o RMSE foi de 0,305 o que é considerado como insatisfatório. Nesse caso foi utilizada a abordagem de cascateamento de modelos de classificação e regressão, após essa ação o RMSE passou para 0,2233 e as predições ficaram dentro de um limite aceitável.

A segunda questão de pesquisa foi: Qual o efeito das janelas de memória para os sistemas preditivos usando apenas dados históricos?

Em resposta ao segundo questionamento, pode se afirmar que para o instrumento INA007 que tem suas medições longe das cotas de tubo e fundo, teve a melhor configuração de acerto em RMSE com o maior número de dias de janela e com o menor número de árvores, aumentando se o número de árvores percebeu-se que o modelo generaliza os resultados.

Para o INA001 que teve suas medições muito próximas a cota de fundo, pode se afirmar que o melhor resultado foi o que utilizou a menor janela de tempo, que no caso foram 2 dias e o número maior de árvores, que nesse caso foram de 100. O modelo com o aumento da janela generaliza, pois, o mesmo tem a tendência de decorar os resultados.

Chega se a conclusão de que, para instrumentos que tem seus dados de medições longe das cotas de fundo e topo, o melhor resultado é aumentando a janela de observação e menor número de árvores no RF, para os instrumentos que possuem os dados muito próximos a cota de fundo, os melhores resultados se dão diminuindo a janela de observações e aumentando o número de árvores do RF.

A terceira questão de pesquisa foi: Existe como avaliar sistema de classificação para o caso do indicador de nível de água fora do limite de leitura do sensor?

Em resposta ao terceiro questionamento, foi constatado através da pesquisa que, não é necessário utilizar o sistema de classificação para o instrumento que apresenta medições longe das cotas de tubo e fundo, pois, a regressão apenas com os valores do próprio instrumento se demonstraram muito satisfatórias, o que pode ser comprovado pelo baixo RMSE do INA007 e análise gráfica entre os valores reais versus os valores preditos. No caso do INA001 o sistema de classificação foi determinante para melhoria dos resultados utilizando o cascateamento do modelo classificador e regressor.

A quarta questão de pesquisa foi: Quais resultados o sistema retorna com inclusão de dados de pluviometria?

Foi constatado na pesquisa que os dados de pluviometria são muito importantes quando o instrumento tem suas medições próximas a cota de fundo e quando há uma moderada correlação de *Pearson* positiva, isso foi constatado para o INA001, pois, com a inserção desses dados pluviométricos houve uma significativa melhora na predição dos valores do nível do lençol freático.

O objetivo geral desse trabalho foi cumprido de forma integral, visto que foi comprovado que é possível realizar previsões de futuras medições dos valores dos indicadores de nível de água, mediante o aprendizado de máquina, com a finalidade de comparar os valores preditos com os valores da carta de riscos. Essa ferramenta pode ajudar para uma atuação preventiva de ações a serem diligenciadas na estrutura no sentido de garantir sua segurança e estabilidade.

Ficou demonstrado que cada sensor terá uma janela de análise e quantidade de árvores no algoritmo *random forest* para os melhores resultados, isso se dá pelo fato do ponto de instalação de cada sensor, pois, as soluções de contorno variam de caso a caso.

Os objetivos específicos do trabalho também foram alcançados, pois, os modelos de regressão e classificação apresentaram resultados sustentáveis e robustos que abonam a criação de um protótipo a ser implementado e validado, podendo o mesmo tornar-se um produto de previsibilidade muito útil para a área de geotecnia de barragens de rejeitos.

# 6. Trabalhos Futuros

Para trabalhos futuros é indicado a utilização de análise multivariada de forma a correlacionar outros tipos de instrumentos geotécnicos, é também indicada a utilização da engenharia de atributos e protocolo de escolha de dados de teste do modelo.

É recomendado a utilização de mais instrumentos que meçam outras grandezas como piezometria, inclinação, recalque para que se possa ter uma abordagem mais holística da instrumentação geotécnica.

Recomenda se que as janelas de observações possam ser variadas com valores diferentes dos utilizados nesta pesquisa, para obter os melhores resultados sem a aplicação de ruídos no modelo.

É indicado que se utilize a previsão de chuva que irá ocorrer nos próximos dias através de canais como o *Weather Channel* (https://weather.com) ou Clima Tempo (www.climatempo.com.br), no sentido de alimentar os modelos de ML e aferir se na data indicada pelos canais, ocorreu de fato a chuva prevista, aferindo também se a predição dos valores dos instrumentos que estão correlacionados com a estabilidade da barragem de rejeitos acertou de forma aceitável as leituras da instrumentação geotécnica.

É de grande relevância estudar uma arquitetura onde os dados da instrumentação geotécnica sejam adquiridos com conexão direta ao banco de dados, tornando a solução completamente *online*, no estudo dessa pesquisa os dados tiveram que ser extraídos de forma *offline*.

## Referências Bibliográficas


ALABOZ, P., DENGIZ, O. DEMIR, S., ŞENOL, H., **Digital mapping of soil erodibility factors based on decision tree using geostatistical approaches in terrestrial ecosystem**, Volume 207, 2021, https://doi.org/10.1016/j.catena.2021.105634.

ALMALAQ, A., **A review of deep learning methods applied on load forecasting**, IEEE International Conference on Machine Learning and Applications (ICMLA), 2017. p. 511-516.

ANALYTICS VIDHYA, **Decision Tree vs. Random Forest**, Disponível em: https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/, Acesso em 10/09/2020.

ARIA, M., CUCURRULO, C., GNASSO, A., **A comparison among interpretative proposals for Random Forests**, https://doi.org/10.1016/j.mlwa.2021.100094, 2021.

BAUER, E., KOHAVI, R., **An empirical comparsion of voting classification algorithms: Bagging, boosting and variants. Machine Learning.** p 36 (1-2), 105-19, 1999.

BAYKASOGLU, A., ÇEVI, A., ÖZBAKIR, L., KULLUK, S., **Generating prediction rules for liquefaction through data mining**, https://doi.org/10.1016/j.eswa.2009.04.033, 2009.

BEDOYA, C., ISAZA, C., DAZA, J., LOPEZ, D., **Automatic identification of rainfall in acoustic recordings**, Ecological Indicators, Volume, p. 95-100 75, 2017.

BISHOP, C. M., **Pattern Recognition and Machine Learning**, Principal component Analysis, p 561-566, Cambridge 2006.

BREIMAN, L**., Bagging predictors**, https://doi.org/10.1023/A:1010933404324, 2001.

CERQUEIRA, H, M, L., **Critérios de projeto para instrumentação piezométrica de diversas estruturas geotécnicas em mineração**. 166 f. Dissertação (Mestrado em Engenharia Geotécnica) – Escola de Minas, Universidade Federal de Ouro Preto, Ouro Preto, 2017.

CRIMINISI, A., SHOTTON, J., **Decision forests for computer vision and medical image analysis**. Springer Science & Business Media. 13, 22, 25, 2013.

CRUZ, P.T., **100 barragens brasileiras: casos históricos, materiais de construção**, projeto, p 82-86, São Paulo, 1996.

DEMŠAR J., ZUPAN B., LEBAN G., CURK T. **Orange: From Experimental Machine Learning to Interactive Data Mining.** In: Boulicaut JF., Esposito F., Giannotti F., Pedreschi D. (eds) Knowledge Discovery in Databases: PKDD 2004. PKDD 2004. Lecture Notes in Computer Science, vol 3202. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30116-5_58, 2004.

EARTH, Google. Disponível em: https://www.google.com.br/intl/pt-BR/earth/, Acesso em 10/07/2021.

EDELMANN, D., MÓRI, T., SZÉKELY G., **On relationships between the Pearson and the distance correlation coefficients**, https://doi.org/10.1016/j.spl.2020.108960, Volume 169, 2021.

FREUND, Y., SHAPIRE, R., **A Decision-Theoretic Generalization of On-Line Learning and an application to Boosting,** J. of Computer ans System Scienes 55, 1997: 119-139.

FREUND, Y., SHAPIRE, R., **A short introduction to boosting. Journal of Japanese Society for Artificial Intelligence,** Vol.14, No5., 1999: 771-780.

FURQUIM, G.; FILHO, G.P.R.; JALALI, R.; PESSIN, G.; PAZZI, R.W.; UEYAMA, J. **How to Improve Fault Tolerance in Disaster Predictions: A Case Study about Flash Floods Using IoT, ML and Real Data.,** https://doi.org/10.3390/s18030907, 2018.

FURQUIM, G., PESSIN, G., FAIÇAL, B., MENDIONDO, E., UEYAMA, J., **Improving the accuracy of a flood forecasting model by means of ML and chaos theory**, https://doi.org/10.1007/s00521-015-1930-z, 2016.

FUSARO. T. C.; **Estabelecimento estatístico de valores de controle para a instrumentação de barragens de terra : estudo de caso das barragens de Emborcação e Piau.** 2007. 155 f. Dissertação (Mestrado em Engenharia Geotécnica) - Universidade Federal de Ouro Preto, Ouro Preto, 2007.

IBRAM, **Gestão e Manejo de Rejeitos de Mineração**, Belo Horizonte, 2016.

GEOKON, **Data sheet** Disponível em: https://www.geokon.com/content/datasheets/2400_Water_Level_Meter_Solinst_101.pdf, Acesso em: 12/07/2020.

HAYKIN, S., **Neural Networks – A Compreensive foundation – 2 edition**, **Neurocomputing**, v. 43, n. 1-4, p. 51-75, New Jersey, 1999.

ICOLD, **Monitoring of Dams and their Foundations**, Bulletin 68, 375p, 1989.

INSTITUTO MINERE, Disponível em: **https://institutominere.com.br/blog/seguranca-de-barragem-perda-estabilidade-talude**, Acessado em 10/06/2021.

INTELLTECH, Disponível em: **https://intelltech.com.br/language/pt/conheca-plataforma-shms/**, Acessado em 13/07/2021.

IPEA, **Diagnóstico dos Resíduos Sólidos da Atividade de Mineração de Substâncias Não Energéticas,** A disposição de rejeitos da mineração, p 20-25, Brasília, 2012.

ITV, **Arquivo interno**, Ouro Preto, 2020.

KHALIFAH, H., GLOVER P.W.J., LORINCZI P., **Permeability prediction and diagenesis in tight carbonates using ML techniques**, https://doi.org/10.1016/j.marpetgeo.2019.104096, 2020.

LEE, S, H. CHOI, H., CHA, K, CHUNG, H, **Random Forest as a potential multivariate method for near-infrared (NIR) spectroscopic analysis of complex mixture samples: Gasoline and naphtha, Microchem**. J. 110 739–748. doi:10.1016/j.microc.2013.08.007, 2013.

MALIAR, L., SERGUEI M., **Solving nonlinear dynamic stochastic models: an algorithm computing value function by simulations**, Economics Letters, Volume 87, Issue 1, 2005.

MAROCO, J., **Análise Estatística – Com utilização do SPSS, 2ª edição**; Edições Sílabo; p 23-29, 2003.

MENGA, R., MASERA, A., BECOCCI, L., JULIANI, M. **Gestão, Tratamento e Interpretação de Dados de Monitoração Estrutural para Controle de Barragens**. XXIII Seminário Nacional de Grandes Barragens, Belo Horizonte, p-30-37, 1999.

MINISTÉRIO DA INTEGRAÇÃO NACIONAL (MI). Secretaria de Infraestrutura Hídrica. **Manual de Segurança e Inspeção de Barragens**. Brasília, p 59-67 2002.

MOHAMED A. S**, State-of-the-art review of some artificial intelligence applications in pile foundations**, https://doi.org/10.1016/j.gsf.2014.10.00, 2016.

PURI, N., PRASAD, H. D., JAIN, A., **Prediction of Geotechnical Parameters Using ML Technique**, https://doi.org/10.1016/j.procs.2017.12.066, 2018.

PRASAD, A, M., IVERSON, L, R., LIAW, A., **Newer classification and regression tree techniques: Bagging and random forests for ecological prediction, Ecosystems**. 181–199. doi:10.1007/s10021-005-0054-1, 2006.

QGIS, Disponível em: **https://www.qgis.org/pt_BR/site/**, Acesso em: 12/10/2021.

ROCSCIENCE, Disponível em: **https://www.rocscience.com/**, 12/10/2021.

SIGMA SENSORS., Disponível em: https://sigmasensors.com.br/produtos/pluviometro-de-bascula-tr-525m, Acesso em: 12/08/2021.

SILVEIRA, J. F. A., **Instrumentação e segurança de barragens de terra e enrocamento**, São Paulo, p 30-34, 2006.

SOARES, L. **Barragem de Rejeitos. Colaboração técnica para o livro Tratamento de Minérios**. Edição 5 - Capítulo 19 – pág. 831-896. Rio de Janeiro: Centro de Tecnologia Mineral (CETEM), 2010.

UÉLISON, J.L.S., PESSIN, G., COSTA, A.C., RIGHI, R.R., **AgriPrediction: A proactive internet of things model to anticipate problems and improve production in agricultural crops**, https://doi.org/10.1016/j.compag.2018.10.010, 2019.

UEYAMA, J., FAIÇAL, B. S., MANOA, L.Y., BAYER, G., PESSIN, G., GOMES, P.H., **Enhancing reliability in Wireless Sensor Networks for adaptive river monitoring systems: Reflections on their long-term deployment in Brazil**, https://doi.org/10.1016/j.compenvurbsys.2017.05.001, 2017.

VALE, Arquivo Interno, Nova Lima, 2018.

VALE, **GEOTEC**, Nova Lima, 2021.

VIEIRA, A, C., GARCIA, G., PABÓN, R.E.C., COTA L. P., SOUZA, P., UEYAMA, J., PESSIN, P., **Improving flood forecasting through feature selection by a genetic**

**algorithm – experiments based on real data from an Amazon rainforest river**, https://doi.org/10.1007/s12145-020-00528-8, 2020.

SVETNIK, V., LIAW, A., TONG, C., CULBERSON, J. C., SHERIDAN, R.P., FEUSTON, B.P., **Random Forest: A Classification and regression tool for compound classification and QSAR modeling**, https://doi:10.1021/ci034160g, 2003.

VOOO, Disponível em: https://www.vooo.pro/insights/um-tutorial-completo-sobre-a-modelagem-baseada-em-tree-arvore-do-zero-em-r-python/, Acesso em 15/08/2021.

XU, S., NIU, R., **Displacement prediction of Baijiabao landslide based on empirical mode decomposition and long short-term memory neural network in Three Gorges area, China**, https://doi.org/10.1016/j.cageo.2017.10.013, 2018.

ZHANG, L.L., ZHANG, J., ZHANG, L.M., TANG, W.H., 2011. **Stability analysis of rainfall induced slope failure**, https://doi.org/10.1680/geng.2011.164.5.299, 2015.

WANG, W, HUANG, M., NGUYEN, Q., HUANG, W., Zhang, K., HUANG, T.,

**Enabling decision trend analysis with interactive scatter plot matrices visualization**,

Journal of Visual Languages & Computing, Volume 33, 2016.