INSTITUTO TECNOLÓGICO VALE

Programa de Pós Graduação em Instrumentação, Controle e Automação de Processos de Mineração - PROFICAM
Universidade Federal de Ouro Preto - Escola de Minas
Associação Instituto Tecnológico Vale - ITV

Dissertação

SENSORES VIRTUAIS BASEADOS EM APRENDIZADO DE MÁQUINA PARA ESTIMATIVA DE VALORES DE SENSORES EM INSTRUMENTAÇÃO BÁSICA DE BARRAGENS DE REJEITO

Bruno Monteiro

Ouro Preto
Outubro de 2023

Bruno Monteiro

SENSORES VIRTUAIS BASEADOS EM APRENDIZADO DE MÁQUINA PARA ESTIMATIVA DE VALORES DE SENSORES EM INSTRUMENTAÇÃO BÁSICA DE BARRAGENS DE REJEITO

Dissertação apresentado ao curso de Mestrado Profissional 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.

Linha de Pesquisa: Tecnologias da Informação, Comunicação e Automação Industrial

Orientador: Prof. D.Sc. Gustavo Pessin
Coorientador: Prof. Ph.D. Frederico Gadelha Guimarães

Ouro Preto, MG – Brasil
Outubro de 2023

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

M775s Monteiro, Bruno Oliveira.
Sensores virtuais baseados em aprendizado de máquina para estimativa de valores de sensores em instrumentação básica de barragens de rejeito. [manuscrito] / Bruno Oliveira Monteiro. - 2023.
74 f.

Orientador: Prof. Dr. Gustavo Pessin.
Coorientador: Prof. Dr. Frederico Gadelha Guimarães.
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 e açudes. 2. Plant Information Management System (PIMS). 3. Aprendizado do computador. 4. Processo decisório. 5. Aprendizado do computador - Gradient Boosting. 6. Detectores. I. Pessin, Gustavo. II. Guimarães, Frederico Gadelha. III. Universidade Federal de Ouro Preto. IV. Título.
CDU 681.5:622.2

Bibliotecário(a) Responsável: Maristela Sanches Lima Mesquita - CRB-1716

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

Bruno Oliveira Monteiro

Sensores virtuais baseados em aprendizado de máquina para estimativa de valores de sensores em instrumentação básica de barragens de rejeito

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 11 de outubro de 2023

Membros da banca

Doutor - Gustavo Pessin - Orientador - Instituto Tecnológico Vale
Coorientador - Frederico gadelha Guimarães - Universidade Federal de Minas Gerais
Membro interno - Juan Manuel Girao Sotomayor - Instituto Tecnológico Vale
Membro externo - Caetano Mazzoni Ranieri - Universidade de São Paulo

Gustavo Pessin, orientador do trabalho, aprovou a versão final e autorizou seu depósito no Repositório Institucional da UFOP em 01/12/2023

Documento assinado eletronicamente por **Bruno Nazário Coelho**, **COORDENADOR(A) DE CURSO DE PÓS-GRADUAÇÃO EM INST. CONTROLE AUTOMAÇÃO DE PROCESSOS DE MINERAÇÃO**, em 14/12/2023, às 11:02, 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 **0641931** e o código CRC **9098BA18**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3552-7352 - www.ufop.br

A alguém cujo valor é digno
desta dedicatória.

Agradecimentos

Caros familiares e estimado professor, gostaria de aproveitar este momento para expressar minha mais profunda gratidão a cada um de vocês. À medida que finalizo minha dissertação, percebo o imensurável valor que cada um trouxe à minha jornada acadêmica. Aos meus amados familiares, meu coração transborda de gratidão por seu apoio incondicional ao longo de todos esses anos. Suas palavras de encorajamento, paciência e amor constante foram fundamentais para que eu persistisse em busca do conhecimento. Cada gesto de carinho, cada palavra de incentivo nos momentos de dúvida, foram a força motriz que me impulsionou a avançar. Sei que posso contar com vocês em todas as fases da minha vida, e isso é um presente inestimável.

Ao meu estimado orientador Gustavo Pessin, desejo expressar minha mais profunda gratidão por sua dedicação e orientação ao longo desta jornada acadêmica. Seu compromisso em transmitir conhecimento, estimular meu pensamento crítico e desafiador e incentivar minha busca por excelência foram verdadeiramente inspiradores. Sua paixão pelo ensino e sua habilidade em cativar a atenção dos alunos são admiráveis. Sou grato por cada minuto dedicado a esclarecer minhas dúvidas, guiar meus projetos e despertar em mim a curiosidade intelectual. Você é um modelo excepcional de professor e mentor.

Nesta oportunidade, também expresso minha gratidão ao Grande Arquiteto do Universo que é Deus. Reconheço sua presença em minha jornada, iluminando meu caminho e oferecendo-me força e sabedoria para enfrentar os desafios. Sua orientação divina permeou cada passo deste percurso acadêmico, e sou humildemente grato por sua influência. A ti, agradeço sinceramente por tornar possível a realização deste trabalho. S.'.F.'.U

Dissertação apresentada à Escola de Minas/UFOP e ao ITV como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

SENSORES VIRTUAIS BASEADOS EM APRENDIZADO DE MÁQUINA PARA ESTIMATIVA DE VALORES DE SENSORES EM INSTRUMENTAÇÃO BÁSICA DE BARRAGENS DE REJEITO

Bruno Monteiro

Outubro/2023

Orientadores: Gustavo Pessin
Frederico Gadelha Guimarães

A aplicação de sensores virtuais na estimativa de dados de instrumentação básica em barragens de rejeito de minério é uma inovação no cenário da mineração atual. Essa abordagem utiliza modelos computacionais avançados para aprimorar o monitoramento dessas estruturas, melhorando a segurança e a eficiência operacional, ao mesmo tempo em que reduz custos. Em um momento em que a gestão responsável dos rejeitos de mineração é vital, os sensores virtuais desempenham um papel fundamental na mitigação de riscos ambientais e na proteção das comunidades próximas às barragens. Dentro do setor de Mineração, o monitoramento de barragens de rejeito tem ganhando bastante notoriedade devido aos últimos incidentes de rompimento ocorridos no Brasil. Nesse aspecto esse trabalho propõe implementar e avaliar métodos de aprendizado de máquina para estimativa de valores para sensores de instrumentação básica utilizados no controle e monitoramento de barragens de mineração.

Palavras-chave: Barragens, Detecção de Anomalias, PIMS, Aprendizado de Máquina, MLP, Random Forest, Decision Tree, Gradient Boosting, Sensores Virtuais
Tema: Sensoreamento de Ativos
Macrotema: Mina
Linha de Pesquisa: Tecnologia da Informação, Comunicação e Automação Industrial
Área relacionada da Vale: Instrumentação básica de barragens

Abstract of research project that will be presented to Escola de Minas/UFOP and ITV as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

MACHINE LEARNING-BASED SOFT SENSORS FOR ESTIMATING SENSOR VALUES IN BASIC INSTRUMENTATION OF ORE TAILINGS DAMS

Bruno Monteiro

October/2023

Advisors: Gustavo Pessin
Frederico Gadelha Guimarães

The application of soft sensors in estimating basic instrumentation data in ore tailings dams is an innovation in the current mining scenario. This approach uses advanced computational models to enhance the monitoring of these structures, improving safety and operational efficiency while reducing costs. At a time when responsible management of mining waste is crucial, virtual sensors play a fundamental role in mitigating environmental risks and protecting communities near the dams. Within the mining sector, the monitoring of tailings dams has gained significant prominence due to recent dam failure incidents in Brazil. In this regard, this work aims to implement and evaluate machine learning methods for estimating values for basic instrumentation sensors used in the control and monitoring of mining dams.

Keywords: Dams, Anomaly Detection, PIMS, Machine Learning, MLP, Random Forest, Decision Tree, Gradient Boosting, Soft Sensors
Theme: Asset Sensing
Macrotheme: Mine
Research Line: Information Technology, Communication and Industrial Automation
Related Area of the Vale: Basic Dam Instrumentation

Lista de Figuras

Lista de Abreviaturas e Siglas

AF Asset Framework

DA Data Archive

DT Decision Tree

LR Linear Regression

MES Manufacturing Execution Systems

PIMS Plant Information Management System

PLC Controlador Lógico Programável (Programmable Logic Controller)

RF Random Forest

SCADA Supervisão Constrole e Aquisição de Dados (Supervisory Control and Data Acquisition)

SDCD Sistema Digital de Controle Distribuído

SDK Kit de Desenvolvimento de Software (Software Development Kit)

Lista de Tabelas

Sumário

Capítulo 1

Introdução

O monitoramento de barragens de rejeito gera um volume imenso de dados, fornecidos por sensores de instrumentação básica, esses dados por sua vez são persistidos em um historiador de dados, contido na camada PIMS (Plant Information Management System) da pirâmide de automação. Essa camada histórica é responsável por armanezar e persistir anos de informações fornecidas por diversos instrumentos contidos nas estruturas de contenção. Os principais instrumentos utilizados para este fim estão listados abaixo:

- Piezômetros: que monitoram por intermédio da poropressão os níveis de água pelo corpo das barragens, assim como em suas respectivas fundações;
- Medidores de nível de água: responsáveis pela medição da água em profundidade nas estruturas, através da cota superficial;
- Medidores de vazão: responsáveis pela medição do fluxo de água que se processa pelo corpo das barragens.

Esses sensores são projetados para compensar perturbações e manter a coleta de informações de forma confiável, entretanto muitas vezes devido a condições extremas podem ocorrer falhas, ou perda de dados, sendo essencial que esses eventos sejam detectados e diagnosticados, para uma correta interpretação da informação colhida. Segundo Soares (2010) a instrumentação montada nos maciços das barragens de rejeitos e em suas fundações tem por fim a segurança estrutural e ambiental da barragem. A instrumentação associada ao controle da segurança ambiental é, basicamente, a mesma utilizada para o acompanhamento do comportamento das barragens convencionais de terra, considerando os aspectos peculiares dos projetos, dos métodos construtivos, dos materiais a serem utilizados no alteamento do maciço e aqueles lançados na bacia de acumulação dos rejeitos. No período operacional da barragem, que se dá durante e após o enchimento do reservatório, a instrumentação tem como objetivo as seguintes operações:

- levantamento de ocorrências eventuais de anomalias que possam colocar em risco a estrutura de contenção;

- verificações de conformidade dos critérios mínimos de operação;

1.1 Contextualização

As barragens de forma geral são construções que tem como objetivo a retenção de materiais sólidos ou líquidos. Na mineração, a barragem de rejeito é utilizada para armazenar todos os materiais não utilizáveis gerados pelo processo de beneficiamento do minério.

Geralmente existe no rejeito de mineração uma elevada quantidade de água, uma vez que a água é bastante utilizada no processo de beneficiamento mineral. As barragens de rejeito desempenham um papel fundamental na indústria de mineração, sendo essenciais para armazenar os resíduos resultantes do processo de extração mineral. A integridade e segurança dessas estruturas são de extrema importância, uma vez que qualquer falha pode resultar em impactos ambientais devastadores e ameaçar vidas humanas. Para garantir a estabilidade e monitorar continuamente o comportamento dessas barragens, a instrumentação adequada desempenha um papel crucial.

A instrumentação de barragens de rejeito envolve a coleta e análise de uma ampla gama de dados, incluindo informações de sensores que medem parâmetros como nível de água, poropressão, temperatura do solo e outros fatores geotécnicos. Tradicionalmente, os dados desses sensores têm sido coletados e interpretados por meio de sistemas de monitoramento convencionais. No entanto, avanços recentes na área de aprendizado de máquina abriram portas para abordagens inovadoras e mais eficazes na interpretação e análise desses dados.

Nesse contexto, surgem os "Sensores Virtuais Baseados em Aprendizado de Máquina". Essa abordagem inovadora aproveita os princípios do aprendizado de máquina para criar modelos preditivos capazes de estimar os valores dos sensores com base em outras variáveis medidas. Em vez de depender exclusivamente dos dados brutos dos sensores individuais, os sensores virtuais podem considerar relações complexas e não lineares entre várias variáveis, permitindo uma estimativa mais precisa e adaptativa dos parâmetros monitorados.

O objetivo desta dissertação é explorar a aplicação de sensores virtuais baseados em aprendizado de máquina na instrumentação básica de barragens de rejeito. Pretende-se demonstrar como esses sensores virtuais podem superar as limitações dos métodos tradicionais de monitoramento, oferecendo uma abordagem mais robusta e precisa para a estimativa de valores de sensores críticos. Além disso, a dissertação irá abordar os benefícios em termos de eficiência operacional, custo e confiabilidade que podem ser alcançados por meio da implementação dessa abordagem inovadora.

Ao longo desta pesquisa, serão analisados exemplos reais de aplicação de sensores virtuais em barragens de rejeito, destacando casos de sucesso e lições aprendidas. Além disso, serão explorados os desafios técnicos e práticos associados à implementação desses sistemas, incluindo questões de coleta de dados, treinamento de modelos de aprendizado

de máquina e validação dos resultados.

1.2 Motivações e justificativas

Segundo Soares (2010) a crescente demanda mundial por bens minerais, aliada ao desenvolvimento econômico e tecnológico, condiciona, de forma sustentável e economicamente viável, o aproveitamento de minérios de baixo teor ou mesmo aqueles de difícil beneficiamento. Esta situação conduz a um aumento expressivo na quantidade de rejeitos produzidos, superando, em muito, aquela advinda dos próprios minérios. Como consequência o aumento na geração de rejeitos tem conduzido a um aumento proporcional das estruturas armazenadoras, fazendo com que, atualmente, as barragens de rejeitos encontrem-se entre as importantes obras da mineração. Concomitantemente ao aumento das dimensões dessas barragens, os vários acidentes ocorridos com as mesmas despertam a atenção da comunidade técnico-científica e de autoridades governamentais para a questão de segurança destas obras. A tabela histórica abaixo ilustra alguns acidentes ocorridos no Brasil e no mundo envolvendo o rompimento de barragens:

Tabela 1.1: Acidentes envolvendo ruptura de barragens

Ano	**Local**	**Causa atribuída**	**Danos Reportados**
1965	El Cobre - Chile	Terremoto/liquefação	210 vítimas, soterramento do povoado.
1970	Mufaline Mine/África	Não definida	89 vítimas – 453.000 m3 de rejeitos saturados
1972	Buffalo Creek/West - Virginia	Não definida	110 mortos, 1100 feridos, 1500 cadas destruídas - 59500m3 de lama.
1974	Impala Platinum - Africa do Sul	Entubamento (piping)	12 vítimas, 3 milhões m3 de lama fluíram por 45 km, destruindo estradas, pontes e soterrando reservatório de água potável.
1985	Prealpi/Trento - Itália	Material de construção	o Liberação de 200.000 m3 de rejeitos. 268 vítimas.
1985	Cerro Negro Chile	Sismo induzido e liquefação	Lama dos rejeitos fluiu até 85 km a jusante.
1985	Pico S.Luiz/MG	Solapamento do pé do aterro e entubamento	Lama fluiu até 10 km a jusante. Pontes e estradas de ferro.
1986	Fernandinho Itaminas - MG	Liquefação	4 vítimas. Destruição de laboratórios e equipamentos.
1996	Mina do Porco/- Bolívia	Entubamento (piping)	3 vítimas – Fazendas, gado, flora e fauna; 300km de rio contaminados.
2015	Fundão - Mariana - MG	Liquefação	19 vítimas além da destruição de vilarejo, fauna e flora.
2019	Córrego do Feijão - Brumadinho - MG	Liquefação	272 vítimas além da destruição de aréa administrativa, fauna e flora.

Fonte: Soares (2010)

Nesse contexto, o estudo e a aplicação de modelos de sensores virtuais baseados em machine learning para contribuição na operação segura de estruturas de barragem de rejeito, através dos requisitos básicos de segurança pode resultar em uma contribuição técnica relevante para comunidades de interesse relacionadas ao tema, sendo portanto a motivação deste trabalho.

1.3 Objetivos

O objetivo principal desse trabalho é explorar a aplicação de sensores virtuais na estimativa de valores de sensores físicos de instrumentação básica de barragens de rejeito dentro da camada PIMS. Com base no objetivo principal os objetivos específicos abaixo foram estabelecidos:

- Propor, implementar e avaliar métodos de aprendizado de máquina baseado em sensores virtuais para estimativa do valor de poropressão de piezômetros de corda vibrante. Os métodos investigados são Rede Neural MLP, Random Forest, Decision Tree, Linear Regression e Gradient Boosting.
- Propor e implementar um método de seleção de vizinhos por meio de correlação, buscando maior eficiência computacional.
- Implementar métodos de memória na série temporal a fim de avaliar o impacto da memória no aprendizado dos métodos.
- Apresentar o funcionamento do modelo selecionado experimentalmente dentro da camada PIMS (OSISoft/AVEVA PI SYSTEM).

1.4 Organização do texto

Esse trabalho está dividido em 6 capítulos.

- No capítulo 1 são abordadas todas as motivações que levaram a construção desse trabalho, assim como a contextualização e justificativas que levaram ao seu desenvolvimento;
- No capítulo 2 temos a revisão bibliográfica, contendo o embasamento teórico de estruturas de barragens e sensores de instrumentação básica, assim como conceitos relevantes para o compreendimentos dos modelos de aprendizado de máquina avaliados;
- No capítulo 3 temos a discussão de trabalhos relacionados ao tema objeto deste trabalho;
- No capítulo 4 temos a explanação dos métodos utilizados neste trabalho, onde são apresentadas as fontes de dados utilizadas, assim como o detalhamento técnico do trabalho;
- No capítulo 5 temos a apresentação de resultados obtidos através dos modelos utilizadas.

- No capítulo 6 temos a aplicação do modelo final dentro da camada PIMS;
- No capítulo 7 temos a conclusão do trabalho, com a análise final dos estudos realizados;
- No capítulo 8 são apresentadas propostas de trabalhos futuros, oferecendo continuidade na proposto dessa dissertação.

Capítulo 2

Revisão Bibliográfica

2.1 Sensores Virtuais

Fortuna (2007) define Sensores Virtuais como modelos matemáticos ou algoritmos que estimam variáveis importantes de um sistema usando dados de entrada, mesmo quando as medições diretas dessas variáveis não estão disponíveis ou são difíceis de obter de maneira confiável. O uso de sensores virtuais, também conhecidos como soft sensors, para séries temporais na indústria, com foco na detecção de anomalias e preenchimento de gaps de dados, está ganhando cada vez mais importância. Isso ocorre porque a indústria moderna coleta grandes quantidades de dados operacionais, mas muitas vezes esses dados podem estar incompletos, corrompidos ou sujeitos a falhas. Os sensores virtuais oferecem uma abordagem eficaz para lidar com essas questões e melhorar a eficiência dos processos industriais, conforme listado abaixo:

- Detecção de Anomalias: Os sensores virtuais podem ser treinados em dados históricos para aprender o comportamento normal do sistema. Quando as séries temporais atuais são comparadas com esses padrões aprendidos, os sensores virtuais podem identificar desvios significativos que indicam anomalias. Isso permite a detecção precoce de problemas operacionais, como falhas em equipamentos ou comportamentos anormais dos processos.

- Preenchimento de Gaps de Dados: Muitas vezes, os dados coletados em ambientes industriais podem conter falhas temporárias devido a problemas de sensores, interrupções de comunicação ou outros fatores. Os sensores virtuais podem preencher esses gaps usando modelos matemáticos que levam em consideração a correlação entre diferentes variáveis. Isso resulta em conjuntos de dados mais completos e confiáveis, que são cruciais para análises precisas e tomadas de decisão informadas.

Sensores Virtuais normalmente utilizam técnicas de modelagem baseadas em dados para construir estimativas das variáveis desejadas. Isso pode envolver técnicas estatísticas,

como regressão, ou mesmo algoritmos de aprendizado de máquina, como redes neurais ou árvores de decisão. A escolha do método depende das características dos dados e da complexidade do processo industrial. Os sensores virtuais podem ser integrados a sistemas de controle e monitoramento existentes, fornecendo informações sobre as variáveis estimadas. Isso permite uma tomada de decisão mais precisa e reativa, resultando em maior eficiência e qualidade do processo.

2.2 Aprendizado de Máquina

Essa seção, tem como objetivo descrever de forma sucinta os métodos de Machine Learning que serão aplicados na geração de sensores virtuais neste trabalho.

2.2.1 Random Forest

O modelo Random Forest é um algoritmo de aprendizado de máquina baseado em ensemble que combina várias árvores de decisão para realizar tarefas de classificação e regressão. É conhecido por sua eficácia em lidar com conjuntos de dados complexos e grandes, além de ser resistente ao overfitting. Hastie (2009) descreve o Random Forest como um método que combina muitas árvores de decisão, geralmente treinadas de forma independente, em um classificador ou regressor. Já Bishop Nasrabadi (2006) descreve o modelo Random Forest como uma técnica bem-sucedida na qual há a combinação de árvores de decisão individuais por meio de uma estratégia de ensemble.

O Random Forest utiliza a técnica de bootstrap para criar conjuntos de treinamento para cada árvore e, durante a construção das árvores, seleciona aleatoriamente um subconjunto de atributos para considerar em cada divisão. Isso ajuda a reduzir a correlação entre as árvores e promove uma maior diversidade na floresta.

Breiman (2001) apresenta o método Random Forest em detalhes e discute suas vantagens, como a capacidade de lidar com conjuntos de dados de alta dimensionalidade e variáveis correlacionadas. Fernández-Delgado (2014) mostra que o Random Forest frequentemente supera outros métodos de classificação em uma ampla gama de problemas.

Em suma, o modelo Random Forest é um algoritmo de aprendizado de máquina baseado em ensemble que combina várias árvores de decisão. As referências destacam a capacidade do Random Forest de lidar com problemas complexos, evitar overfitting e fornecer resultados robustos em uma variedade de cenários.

2.2.2 Decision Tree

O modelo Decision Tree é um algoritmo de aprendizado de máquina que constrói uma estrutura em forma de árvore para tomar decisões com base em regras e condições definidas pelos dados de treinamento. É uma abordagem intuitiva e interpretável que é amplamente

utilizada em diversas aplicações. Hastie (2009) descreve o modelo como uma estrutura de divisão recursiva que divide os dados em segmentos cada vez menores, de acordo com os valores dos atributos preditivos. Já Bishop Nasrabadi (2006) descreve o as árvores de decisão como uma abordagem popular para a construção de modelos de classificação onde é gerada uma divisão do espaço de entrada em regiões simples e potencialmente descontínuas.

As árvores de decisão são construídas a partir de uma série de decisões binárias com base nos atributos dos dados. Cada nó interno da árvore representa uma condição de divisão e cada folha representa uma classe ou valor de saída. Durante o treinamento, o algoritmo de construção da árvore procura as melhores divisões dos dados, com o objetivo de maximizar a pureza das classes nas folhas ou minimizar o erro de regressão.

Breiman (2017) em seu trabalho introduz o algoritmo CART (Classification and Regression Trees) e descreve sua construção e propriedades relacioando as principais características do modelo.

Em resumo, o modelo Decision Tree é um algoritmo de aprendizado de máquina que utiliza uma estrutura em forma de árvore para tomar decisões com base em regras e condições definidas pelos atributos dos dados.

2.2.3 MLP

O modelo MLP (Multilayer Perceptron) é um tipo de rede neural artificial com múltiplas camadas. Ele é composto por uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Cada camada contém um conjunto de neurônios interconectados que processam informações e passam os resultados para a próxima camada.

Goodfellow (2016) descreve o Multilayer Perceptron como uma rede neural feed-forward composta por várias camadas de neurônios, nas quais os neurônios em uma camada se conectam a neurônios na camada subsequente. O autor também fornece uma descrição sobre as funções de ativação no MLP, na qual é dito que a mesma é aplicada à combinação linear de suas entradas para introduzir não linearidades na rede. Exemplos comuns de funções de ativação incluem a função sigmoide logística, a função tangente hiperbólica e a função ReLU. Em Bishop Nasrabadi (2006) o Multilayer Perceptron é descrito como a forma mais básica de uma rede neural de múltiplas camadas e consiste em uma série de camadas de unidades de processamento interconectadas.

O MLP é conhecido por sua capacidade de aprender representações complexas e realizar tarefas como classificação e regressão. Ele usa funções de ativação não lineares em cada neurônio para introduzir não linearidades na rede. Essas funções, como a função sigmoide ou a função ReLU, ajudam o MLP a aprender relações não lineares nos dados. Através do uso de técnicas como a retropropagação do erro e o gradiente descendente, o MLP é treinado para ajustar os pesos sinápticos das conexões entre os neurônios. Isso

permite que o modelo aprenda a mapear os padrões de entrada para a saída desejada.

2.2.4 Linear Regression

O modelo de Regressão Linear é um dos métodos mais fundamentais e amplamente utilizados no campo da análise estatística e do aprendizado de máquina. Ele busca estabelecer uma relação linear entre uma variável dependente contínua e uma ou mais variáveis independentes. Bishop Nasrabadi (2006) descreve a Regressão linear como um dos métodos mais antigos e amplamente utilizados para a análise de dados. Hastie (2009) aborda a Regressão Linear como um modelo amplamente estudado e comumente utilizado para modelagem preditiva e inferência de dados.

O modelo de Regressão Linear busca estimar os coeficientes que representam a relação linear entre as variáveis independentes e a variável dependente. Isso é feito minimizando a soma dos quadrados dos erros entre as previsões do modelo e os valores reais dos dados de treinamento. Existem diferentes variantes da Regressão Linear, como a Regressão Linear Simples (com uma única variável independente) e a Regressão Linear Múltipla (com múltiplas variáveis independentes). Hastie (2009) discute em detalhes a Regressão Linear e suas extensões, bem como a importância da regularização para lidar com problemas de sobreajuste (overfitting), que costuma ser um efeito colateral característico deste modelo.

2.2.5 Gradient Boosting

O modelo Gradient Boosting é um poderoso algoritmo de aprendizado de máquina que pertence à família de métodos de boosting. Ele combina múltiplos modelos de regressão ou classificação fracos para construir um modelo forte e de alta precisão. Segundo Hastie (2009) o Gradient Boosting é um procedimento geral para a construção de modelos aditivos por estágios, onde os modelos fracos são ajustados em relação aos resíduos dos modelos anteriores.Bishop Nasrabadi (2006) define Gradient Boosting como uma técnica de boosting que se baseia em ajustar um modelo aditivo fraco por meio da minimização de uma função de perda diferenciável.

Em suma, o Gradient Boosting trabalha iterativamente, treinando modelos fracos em sequência e ajustando-os para minimizar a função de perda diferenciável. A cada iteração, o modelo tenta corrigir os erros cometidos pelos modelos anteriores, direcionando a atenção para os exemplos difíceis de serem previstos corretamente. Isso resulta em um modelo final robusto e de alta precisão.

2.3 Séries Temporais

Segundo Ehlers (2007) uma série temporal é definida como uma coleção de observações feitas sequencialmente ao longo do tempo. Sendo portanto a característica mais impor-

tante desse tipo de dados a relação dependente entre observações vizinhas. Geralmente em modelos de regressão a ordem das observações é irrelevante, ao passo que em séries temporáis a ordem dos dados é crucial. As análises vinculadas a séries temporais possuem algumas características inerentes, como as exemplificadas abaixo:

- Possuem informações correlacionadas e consequentemente uma complexidade agregada maior em sua análise;
- A sequência temporal das observações deve sempre ser levada em consideração;
- Geralmente possuem alguns complicadores como a presença de tendências sazonais ou cíclicas que devem ser consideras;
- A detecção de observações anômalas torna-se mais complexa devido à natureza sequencial da informação;
- A seleção do modelo adequado para a análise temporal não é tarefa trivial, uma vez que a natureza da informação está sempre atrelada a características singulares;

2.4 Anomalias

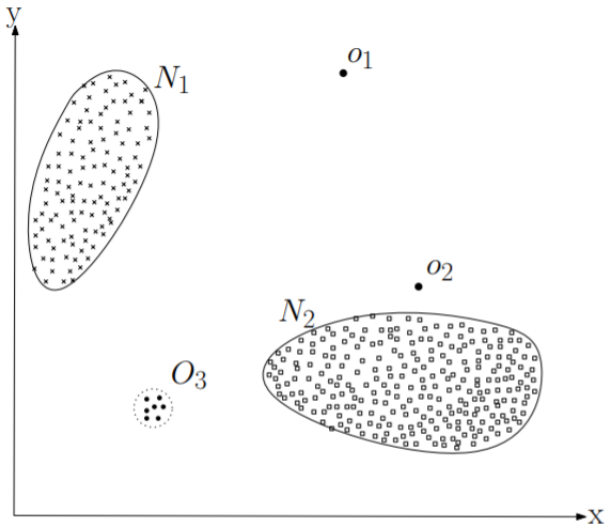
Tem-se como definição de anomalia padrões ou eventos em um conjunto de dados que não correspondem a um conceito bem definido de normalidade Varun Chandola (2009). Ou seja, são dados que demonstram comportamento estranho, diferente de um padrão dito como normal. A Figura abaixo representa um conjunto de dados bidimencional, no qual há duas regiões normais, representadas por N1 E N2. Os pontos que não se aproximam suficientemente dessas regiões, como demonstrado nas marcações O1, O2 e O3 são considerados anômalos.

Nesse contexto existem alguns desafios no problema de detecção de anomalias em um conjunto de dados:

- Durante a modelagem da região de normalidade é extremamente difícil definir fronteiras com características representativas entre normalidade e anormalidade;
- Geralmente as definições de normalidade não são estáticas, podendo portanto sofrer alteração de acordo com modificações no processo de acsição do dados, tornando-se portando não representativas ao longo do tempo;
- Amostras de anormalidade geralmente são raras, tornando o processo de classificação complexo;

Quando inicia-se o processo de análise de anamolias em um conjunto de dados, um dos principais aspectos que deve ser observado é a natureza do dado, uma vez que a

Figura 2.1: Gráfico bidimensional de anomalia em um DataSet



Fonte: Varun Chandola (2009)

técnica a ser utilizada dependerá diretamente das propriedades inerentes ao conjunto de observações. A natureza do dado define a relação de aplicabilidade das diferentes técnicas de detecção de anomalias. Essa grande variabilidade na natureza dos dados cria conceitos distintos de anomalias a serem detectadas na análise da informação, sendo essas divididas nos seguintes tipos:

- **Anomalias Pontuais:** tipo mais simples de anomalia, definida como uma amostra que divergente de todo o restante do conjunto analisado;
- **Anomalias Contextuais:** tipo mais comum em séries temporais de dados. Nesse tipo de anomalia uma amostra é considerada anômala somente em um contexto específico, podendo essa mesma amostra ser considerada normal fora desse contexto.
- **Anomalias Coletivas:** segundo Varun Chandola (2009) uma coleção de dados relacionados é anômala em relação a todo conjunto de dados. Instâncias de dados individuais por si mesmas não podem ser anômalas, mas a ocorrência conjunta como uma coleção é dita anômala.

2.4.1 Técnicas para detecção de anomalias

A literatura geralmente divide as técnicas de deteção de anomalias em três categorias, de acordo com o tipo de informação disponíveis para utilização na modelagem, como descrito abaixo:

- **Detecção não supervisionada:** quando não há necessita de dados de treinamento;

- **Detecção semi-supervisionada:** quando as classes normais já estão definidas. Não há exigência de anomalias rotuladas;
- **Detecção supervisionada:** quando existe a disponibilidade de conjuntos de dados rotulados para as classes normais e anômalas.

Em relação a classificação das observações geralmente utiliza-se uma rotulação para a categorização das instâncias ou uma pontuação para classificação do grau de normalidade ou anormalidade das instâncias contidas na série temporal.

2.5 Instrumentação de barragens de mineração

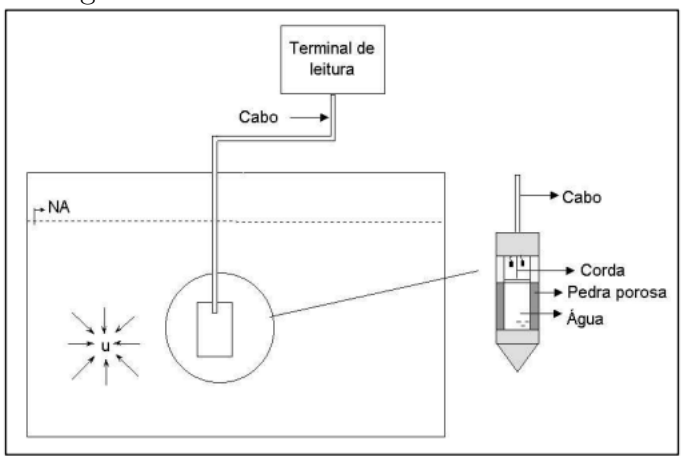
O monitoramento de barragens de rejeito é de suma importância para as empresas de mineração, levando-se em consideração o dano potencial que uma ruptura estrutural pode causar. Durante décadas as barragens de rejeitos foram gerenciadas por intermédio de sistemas manuais de monitoramento, essa atividade envolvia visitas periódicas a pontos pré-selecionados nas estruturas de contenção. Nesses pontos medições manuais eram feitas com alguns instrumentos, como medidores de nível de água, piezômetros e inclinômetros. Entretando esse tipo de medição manual não gera volume de dados suficiente para um monitoramento robusto das barragens. Devido a esse fato, nas décadas mais recentes as mineradoras tem implantado soluções autônomas, na qual os dados são enviados de forma contínua. O monitoramento contínuo é essencial para prevenir falhas nas barragens de rejeito que venham a culminar em sobrecarga, comportamento anômalo da estrutura geotécnica ou até mesmo falha nos mecanimos de drenagem, ocasionando no aumento da poropressão de água e perca da resistência mecânica da estrutura. Desse modo a instrumentação autônoma de barrages é importante para assegurar o constante monitoramento de suas estruturas, obtendo maior segurança e confiabilidade com relação a estabilidade deste empreendimento para cumprir os requisitos legais, segundo Markle Fernandes Vieira (2017).

Os instrumentos instalados em barragens ou em sua fundação podem fornecer informações a respeito do seu comportamento, sendo que em fase do projeto do empreendimento contribuem indicando a possibilidade de adequação e revisão, já durante a operação apontam os possíveis desvios que requerem ações corretivas. Além disso, servem como base de dados e parâmetros para construção de novos projetos segundo Gaioto (2003). O presente trabalho limitou-se à utilização de dados referentes a piezômetros de corda vibrante, entretanto visando dar embasamento para o uso futuro de dados de múltiplos sensores, outros instrumententos foram incorporados à revisão bibliográfica deste trabalho.

2.5.1 Piezômetro de Corda Vibrante

Segundo FONSECA (2003) os piezômetros de corda vibrante são instrumentos constituídos por um corpo cilíndrico de aço inox, alojando internamente uma pedra porosa e uma membrana de aço inox, em cuja face é fixado um fio de aço (corda) tensionado e passando através de um eletro-imã, conforme exibido na Figura 2.2. A blindagem dos cabos elétricos de conexão entre a célula piezométrica e o medidor externo constitui procedimento fundamental para garantir a integridade do instrumento contra efeitos de sobretensões e/ou descargas elétricas.

Figura 2.2: Piezômetro de Corda Vibrante

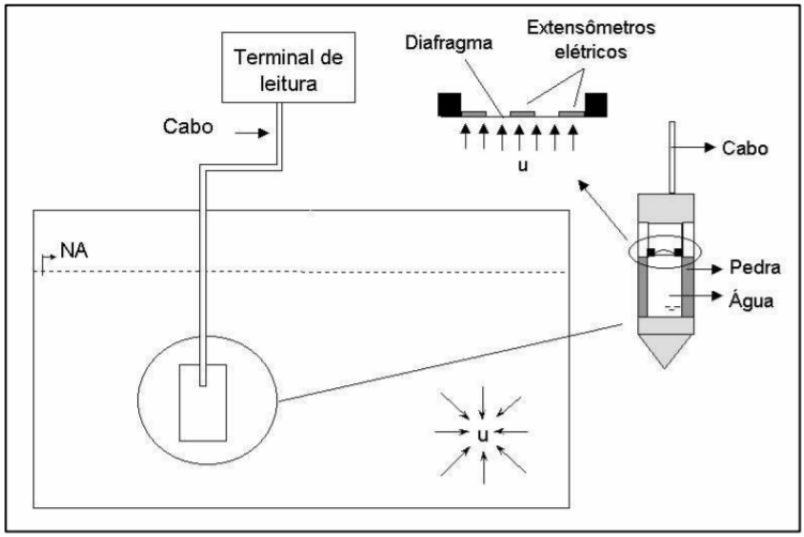


Fonte: FONSECA (2003)

2.5.2 Piezômetro Elétrico

No piezômetro elétrico a poropressão da água de um ponto específico da estrutura da barragem é monitorada por um transdutor. Esse tipo de piezômetro apresenta baixo tempo de resposta e elevada precisão, uma vez que uma pequena alteração no volume da água já gera um movimento no diafragma do transdutor, gerando variação no mesmo. Devido a esse fato esse instrumento é utilizado para obtenção de pressões no maciço de terra, nos taludes e nas fundações. Toda e qualquer pressão externa aplicada ao diafragma produz um pequeno sinal elétrico proporcional a sua movimentação. Dunnicliff (1988) apresenta o método construtivo de piezômetros. Resumidamente, o método consiste na realização de uma perfuração do subsolo com diâmetro de aproximadamente 150 mm. Após a perfuração, insere-se um tubo geomecânico ranhurado com diâmetro de aproximadamente 100 mm, envolvido por uma manta geotêxtil. Após a construção, aguarda-se a estabilização do nível d'água e o instrumento estará pronto para o monitoramento. A Figura abaixo mostra a representação de um piezômetro elétrico:

Figura 2.3: Piezômetro Elétrico

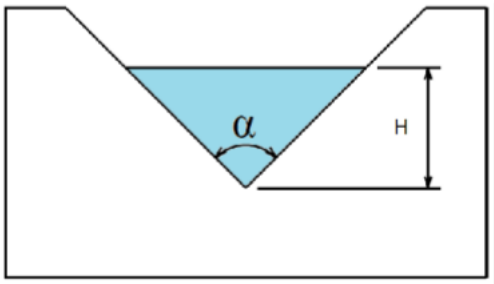


Fonte: FONSECA (2003)

2.5.3 Medidor de vazão

Para a inferência da vazão um procedimento típico é utilizado, segundo Machado (2007), desvia-se o fluxo das barragens para caixas de concreto, nas quais são instalados pequenos vertedouros triangulares ou calhas Parshall. Para as vazões mais reduzidas os vertedouros triangulares permitem uma maior precisão e nesse caso o fluxo de água é direcionado através de uma chapa triangular em forma de "V", com lados ortogonais idênticos. Então mede-se o valor da lâmina de água neste vertedouro, conforme ilustrado na Figura que segue:

Figura 2.4: Medidor de Vazão



Fonte: Autoria Prórpria

Machado (2007) esses medidores de vazão são dispositivos para medição de vazão de canais abertos e a partir da altura da lâmina d'agua na seção convergente é possível encontrar o valor de vazão aproximada pela relação a seguir:

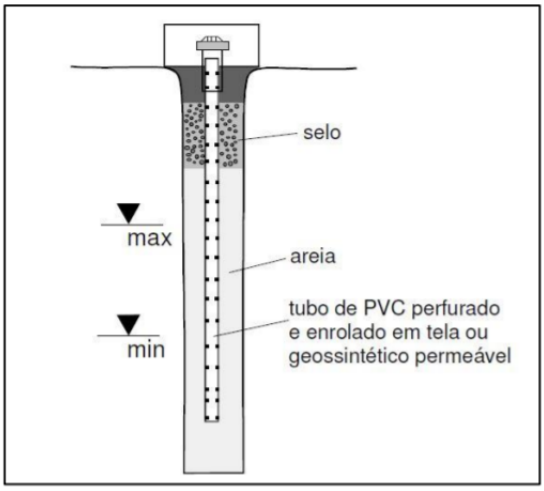
$$Q = 1,4xH^{\frac{5}{2}} \tag{2.1}$$

onde Q é a vazão em m^3/s e H é a altura da lâmina de água em metros.

2.5.4 Indicador de nível de água

Em conformidade com a indicação de Machado (2007) o medidor de nível é o instrumento mais simples de ser construído e operado, tal instrumento tem como objetivo a determinação do nível do lençol freático na estrutura. A configuração desse instrumento é semelhante a do piezômetro, geralmente utiliza-se um tubo em PVC com furos em suas extremidades de modo a permitir a entrada de água. Nesse parte coloca-se uma tampa para estanque, onde um material filtrante é envolvido com uma manta geotêxtil. Após a instalação da estrutura do medidor de nível deve-se realizar a leitura de referência "zero"deste instrumento, para apartir desse ponto basear as demais medições e análises.

Figura 2.5: Medidor de nível



Fonte:FONSECA (2003)

2.6 Coleta de dados

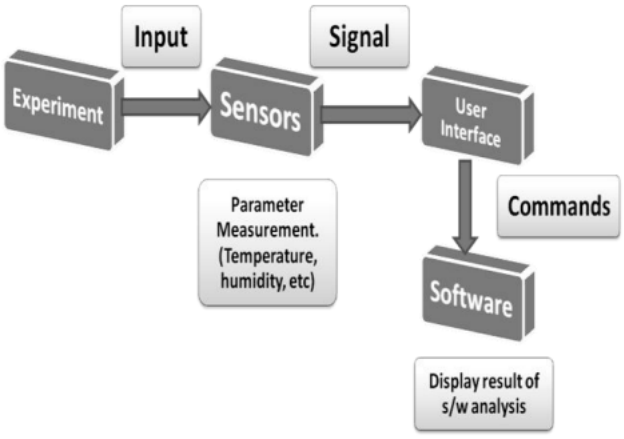
2.6.1 Data Logger

Segundo Badhiye (2011) um Data Logger é um dispositivo que trabalha com sensores para converter fenômenos físicos e estímulos em sinais eletrônicos como voltagem e corrente por exemplo. Esses sinais eletrônicos são convertidos em dados binários, que por sua vez são facilmente interpretados por softwares e armazenados para análise de processos. A característica do Data Logger é levar medições de sensores e armazenar os dados para

uso futuro. No entanto, uma aplicação de registro de dados raramente requer apenas aquisição de dados e armazenamento. A capacidade de analisar e apresentar os dados para determinar resultados e tomar decisões com base nos data logger são necessários. Uma aplicação de data logging completa geralmente requer os elementos ilustrados abaixo:

- **Experiência:** Os vários parâmetros cujos valores devem ser registrados a partir de um ambiente ou objeto específico são fornecidos como entrada para os sensores na parte do experimento;
- **Sensores:** As entradas de várias fontes são dadas ao Data Logger através de sensores para medir parâmetros como temperatura e umidade, para tanto sinais elétricos são convertidos em valores;
- **Interface do usuário:** É fornecida uma interface para interação com o software e os sensores, e usando um algoritmo implementado, a análise é realizada para armazenar os dados;
- **Software:** Ele exibe as informações armazenadas pelos sensores e também mantém os dados para armazenamento a longo prazo.

Figura 2.6: Processo de Data Logging



Fonte: Badhiye (2011)

A Figura 2.7 ilustra um Data Logger da fabricante Campbell, comumente utilizado no monitoramento de barragens de rejeito.

2.6.2 Loggernet

De acordo com Guide (2002) o Loggernet é o principal pacote de software oferecido pela empresa para data logger. Ele suporta programação, comunicação e recuperação de dados

Figura 2.7: DataLogger: CR300

Fonte: Scientific (2013)

entre data loggers e um computador. Os registradores de dados Campbell Scientific processam medições feitas com uma ampla variedade de sensores e armazenar resumos e estatísticas dessas medições como dados. Os registradores de dados realizam essas operações com base em instruções. Programas podem ser enviados para registradores de dados através de uma variedade de canais de comunicação. Uma rede de registradores de dados pode conter de uma a várias centenas de registradores de dados, cada um com seu próprio conjunto de sensores, links de comunicação e cronograma para fazer as medições. O gerenciamento de redes de registradores de dados é uma intrincada tarefa. O software Campbell Scientific LoggerNet, lida com a tarefa de gerenciamento dessa rede, executando as tarefas descritas abaixo:

- Envio de programas criados pelo usuário para registradores de dados;
- Verificação dos tempos do computador e do registrador de dados para sincronismo da informação;
- Gerenciamendo do cronograma de coleta de dados de cada data logger, para definição de periodicidade das consultas;
- Armazenamento de valores no cache de dados do servidor LoggerNet em tabelas para persistência dos dados;
- Manutenção de arquivos de log para todas as comunicações, erros e etc.

2.7 PIMS (Plant Information Management Systems)

Segundo Carvalho (2013) PIMS são sistemas de aquisição de dados que, basicamente, recuperam os dados do processo residentes em fontes distintas, os armazenam num banco de dados único e os disponibilizam através de diversas ferramentas. Apartir de uma estação de trabalho, pode-se visualizar tanto os dados de tempo real como históricos

da planta. Pode-se montar tabelas, gráficos de tendência, telas sinópticas e relatórios dinâmicos, concentrando a informação e possibilitando uma visão unificada de todo o processo produtivo.

O esquema demostrado na Figura 2.8 através da pirâmide da automação industrial tenta organizar os diferentes níveis de controle existentes através da divisão em cinco níveis hierárquicos dos dados. Os níveis mais baixos estão diretamente relacionados com os equipamentos utilizados em campo, enquanto os níveis superiores tratam do gerenciamento dos dados de processos, da planta e da empresa.

- **Sensores e atuadores:** O primeiro nível é majoritariamente composto por dispositivos de campo. Atuadores, sensores, transmissores e outros componentes presentes na planta compõem este nível.

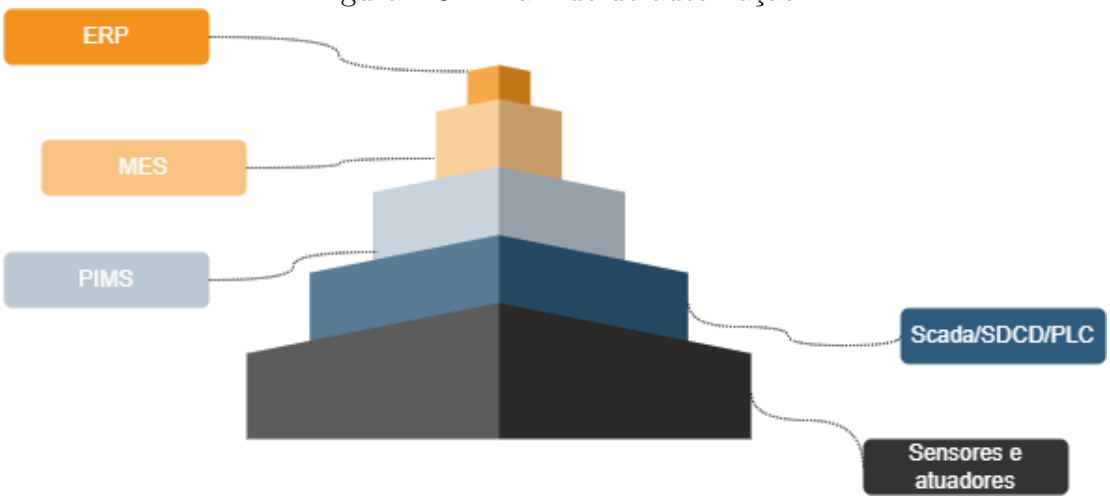
- **SCADA/SDCD/PLC**: O segundo nível compreende equipamentos que realizam o controle automatizado das atividades da planta. Aqui se encontram PLCs (Controlador Lógico Programável), SDCD's (Sistema Digital de Controle Distribuído) e relés.

- **PIMS - Plant Information Management System**: O terceiro nível destina-se a supervisão dos processos executados por uma determinada célula de trabalho em uma planta. Na maioria dos casos, também obtém suporte de um banco de dados com todas as informações relativas ao processo.

- **MES - Manufacturing Execution Systems**: O quarto nível é responsável pela parte de programação e também do planejamento da produção. Auxilia tanto no controle de processos industriais quanto também na logística de suprimentos. Podemos encontrar o termo Gerenciamento da Planta para este nível.

- **ERP - Enterpise Resource Planning**: O quinto e último nível da pirâmide da automação industrial se encarrega da administração dos recursos da empresa, nessa camada é realizado o planejamento estratégico e o gerenciamento corporativo da informação. Neste nível encontram-se softwares para gestão de venda, gestão financeira e BI (Business Intelligence) para ajudar na tomada de decisões que afetam a empresa como um todo.

Figura 2.8: Pirâmide de automação



Fonte: Os Autores

2.7.1 PI System

O PI System é um sistema historiador da camada PIMS disponibilizado pela empresa OSISoft/AVEVA. Este por sua vez coleta, armazena e gerencia dados de qualquer planta ou processo. Com ele os usuários são capazes de se conectar às suas fontes de dados através da coleta dessas informações que são enviadas para o PI Data Archive. O PI Asset Framework (PI AF) oferece uma camada contextualizadora de informações, que permite o acesso mais intuitivo aos dados persistidos anteriormente no PI Data Archive (PI DA). Existe também a possibilidade de se agregar outras informações pertinentes ao processo que não estão atreladas à coleta de informações na automação. Os usuários finais podem por sua vez solicitar dados do PI AF Server ou PI Data Archive para exibição nas ferramentas cliente (PI Vision, PI Data Link e etc) e desta forma trabalhar em análises sobre as séries temporais. Geralmente, as partes envolvidas em um PI System são as descritas abaixo:

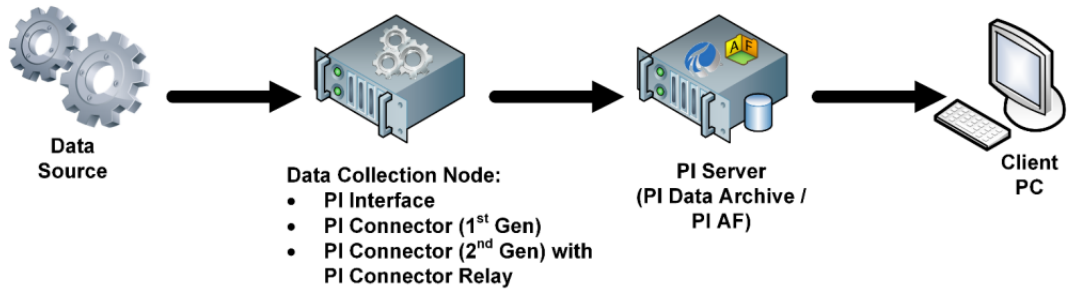
- Data Sources que tem suas informações coletadas por uma PI Interface ou por um PI Connector (contidas em uma máquina de aquisição de dados destinada para esta finalidade);
- PI Interface que tem como função transformar informação coletada em série temporal interpretável para o contexto do PI System e então enviar esse dado diretamente para o PI Data Archive Server;
- PI Connector (primeira geração) que tem como função transformar informação coletada em série temporal interpretável para o contexto do PI System e então enviar esse dado diretamente para o PI Data Archive Server, oferecendo também o recurso

de criação automática dos PI Points, além do envio da estrutura contextualizada do data source para o PI Asset Framework;

- PI Connector (segunda geração) que tem como função transformar informação coletada em série temporal interpretável para o contexto do PI System e então enviar esse dado ao PI Connector Relay que por sua vez encaminha esse pacote para o PI Data Archive e para o PI Asset Framework. Nessa opção todo o gerenciamento do PI Connector (segunda geração) é feito pelo PI Data Collection Manager.

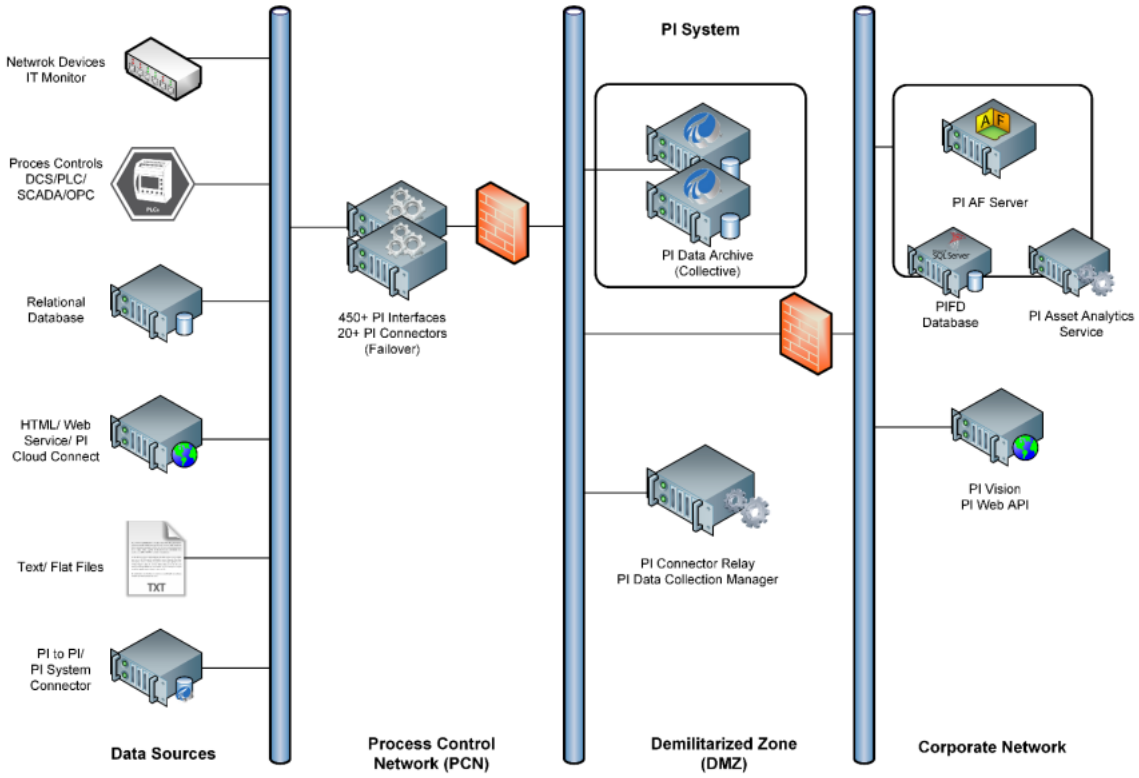
Esse processo está expresso nas Figuras 2.9 e 2.10 que representam respectivamente um PI System padrão em componentes e infraestrutura de rede:

Figura 2.9: PI System: Componentes Básicos



AVEVA (2022)

Figura 2.10: PI System: Arquitetura de Rede Padrão



Fonte: AVEVA (2022)

2.7.2 PI AF SDK

O PI AF SDK é uma biblioteca .NET que provém uma variedade de estruturas de acesso aos dados armazenados no PI System. Esses dados incluem:

- Assets, suas propriedades e relacionamentos;
- Séries temporais contidas no PI Data Archive;
- Event frames que armazenam e contextualizam processos baseados a eventos;

O PI AF SDK foi desenvolvido para utilização através das linguagens da Microsoft .NET, como Visual Basic, C Sharp e C++. Algumas adaptações permitem também sua utilização em outras liguagens de programação, como por exemplo Python (objeto deste trabalho).

Capítulo 3

Trabalhos Relacionados

Müller (2020) fornece uma definição clara de sensores virtuais e sua importância na obtenção de informações de variáveis não medidas diretamente. Os sensores virtuais são modelos matemáticos que estimam essas variáveis com base em informações disponíveis de outras variáveis medidas. O artigo apresenta uma ampla gama de aplicações dos sensores virtuais em diferentes áreas, como indústria química, petroquímica, bioengenharia, sistemas de energia, processos ambientais, entre outros. Exemplos específicos são fornecidos para ilustrar como os sensores virtuais podem ser usados em diferentes contextos. São discutidas diversas técnicas utilizadas na construção de sensores virtuais, incluindo abordagens baseadas em modelos físicos, técnicas de regressão, algoritmos de aprendizado de máquina e métodos de otimização. O artigo explora as vantagens e desafios associados a cada uma dessas abordagens. O artigo aborda também a importância da avaliação de desempenho dos sensores virtuais e apresenta várias métricas e métodos para avaliar a precisão, robustez e confiabilidade desses sensores. São discutidas também as estratégias para calibrar e validar os sensores virtuais em diferentes cenários. O artigo identifica e discute os desafios enfrentados na implementação de sensores virtuais, como a seleção de variáveis de entrada, a modelagem adequada da relação entre as variáveis e o tratamento de incertezas nos dados. São exploradas também as oportunidades futuras na área de sensores virtuais, como a integração com a Internet das Coisas (IoT) e a inteligência artificial. Os autores concluem destacando o potencial dos sensores virtuais como uma solução promissora para obter informações sobre variáveis não medidas em diferentes domínios industriais. Ele enfatiza a importância de uma abordagem integrada, considerando as características específicas de cada aplicação e a seleção adequada das técnicas de modelagem e inferência.

Yan (2022) propõem um método para detecção de anomalias em equipamentos usando uma combinação de sensores virtuais e uma rede de atenção em grafo. O desafio abordado no artigo é a presença de dados incompletos, onde algumas informações dos sensores podem estar ausentes. Para lidar com esse problema, os autores propõem o uso de sensores virtuais, que são capazes de estimar os valores ausentes com base nos dados disponíveis.

O método proposto utiliza uma rede de atenção em gráfico, que modela as interações entre os sensores e captura relações complexas entre eles. Essa rede é treinada para reconhecer padrões normais de operação dos equipamentos e identificar anomalias. O artigo descreve a implementação do método em um conjunto de dados de equipamentos industriais e apresenta experimentos e resultados de desempenho. Os resultados mostram que o método proposto é capaz de detectar anomalias, mesmo quando os dados estão incompletos.

Sobreira (2021) discorre sobre a utilização de sensores virtuais e técnicas de aprendizado de máquina para estimar a vazão mássica de minério de ferro em sistemas de transporte por correias transportadoras. O autor inicia a dissertação apresentando o contexto e a importância da estimativa da vazão mássica em operações de mineração e transporte de minério de ferro. Em seguida, são abordados os desafios associados à medição direta da vazão mássica nesse tipo de sistema, como a dificuldade de instalar sensores físicos convencionais em locais estratégicos ao longo das correias transportadoras.

Para contornar esses desafios, o autor propõe a utilização de sensores virtuais, que são modelos de inteligência computacional capazes de estimar a vazão mássica com base em outras variáveis de processo que podem ser medidas. Em particular, técnicas de árvores de decisão são aplicadas para desenvolver os sensores virtuais, utilizando dados históricos de vazão mássica e outras variáveis disponíveis. O autor descreve a metodologia utilizada para desenvolver e treinar os sensores virtuais, bem como a seleção e o processamento dos dados utilizados no estudo. Os autores apresentam os resultados experimentais obtidos com a aplicação dos sensores virtuais em correias transportadoras reais. São discutidas a acurácia e a eficácia das estimativas de vazão mássica obtidas com o uso dos sensores virtuais, comparando os resultados com medições reais e demonstrando a viabilidade e utilidade dessa abordagem. Ao final do artigo, o autor conclui discutindo as contribuições do trabalho, as limitações encontradas e possíveis direções futuras para aprimorar e expandir a aplicação dos sensores virtuais na estimativa da vazão mássica em correias transportadoras de minério de ferro. Os autores demonstram em Sobreira (2023) uma aplicação prática do trabalho descrito acima, no qual dois sensores virtuais são implementados em um controlador lógico programável de uma correia transportadora em uma área de mineração, onde foi possível verificar o desempenho dos sensores virtuais em uma situação real. Como resultado, os sensores virtuais propostos foram capazes de medir o fluxo de minério com uma taxa de erro aceitável em comparação com uma balança de correia física. A produção estimada dos sensores propostos também se mostrou próxima da produção total real registrada.

Furquim (2018) mostra como o aumento do número da intensidade de desastres naturais é um problema sério que afeta o mundo inteiro. As consequências desses desastres são significativamente piores quando ocorrem em áreas urbanas devido às vítimas e à extensão dos danos a bens e propriedades. Até o momento, os métodos viáveis para lidar

com isso incluíram o uso de redes de sensores sem fio (WSNs) para coleta de dados e técnicas de aprendizado de máquina (ML) para prever desastres naturais. No entanto, os autores mostram como recentemente houveram inovações tecnológicas promissoras que complementaram a tarefa de monitorar o ambiente e realizar previsões. Um desses esquemas envolve a adoção de redes de sensores baseadas em IP (Internet Protocol), usando padrões emergentes para IoT (Internet das Coisas). Diante disso, neste estudo os autores descrevem os resultados alcançados pelo SENDI (Sistema para Detecção e Previsão de Desastres Naturais com base em IoT). O SENDI é um sistema tolerante a falhas baseado em IoT, ML e WSN para a detecção e previsão de desastres naturais e emissão de alertas. O sistema foi modelado por meio do ns-3 e dados coletados por uma WSN do mundo real instalada na cidade de São Carlos, Brasil, que realiza a coleta de dados dos rios da região. A tolerância a falhas está incorporada no sistema, antecipando o risco de interrupções na comunicação e destruição dos nós durante desastres. Ele opera adicionando inteligência aos nós para realizar a distribuição de dados e previsões, mesmo em situações extremas. Um estudo de caso também é incluído para a previsão de enchentes repentinas e isso utiliza o modelo SENDI ns-3 e dados coletados pela WSN.

Em Ueyama (2017) os autores abordam a questão da melhoria da confiabilidade em Redes de Sensores Sem Fio (Wireless Sensor Networks - WSNs) em sistemas de monitoramento de rios adaptativos a longo prazo no Brasil. O estudo está relacionado a sistemas de cidades inteligentes que podem ser adaptados com sucesso para desastres naturais e questões de segurança pública. O artigo é iniciando abordando o crescimento urbano desenfreado e como ele tem levado ao aumento de desastres naturais. Nesse contexto a tecnologia de WSNs tem sido investigada para superar esses problemas, alinhando-se ao conceito de cidades inteligentes incluindo monitoramento de enchentes urbanas. Segundo os autores em ambientes críticos e dinâmicos, a confiabilidade das WSNs é crucial, uma vez que pequenos erros podem ter um grande impacto, especialmente em sistemas que exigem informações em tempo real para tomada de decisões. Nesse aspecto o artigo explora quais são os fatores-chave para fornecer um sistema WSN confiável em sistemas críticos, como monitoramento de enchentes e também como desenvolver um sistema WSN crítico adaptável e confiável.

Em Vieira (2021) os autores abordam um problema de seleção de features com o objetivo de melhorar um modelo de previsão de enchentes. O modelo proposto é desenvolvido por meio de um estudo de caso que utiliza dezoito séries temporais diferentes de trinta e cinco anos de dados hidrológicos, prevendo o nível do Rio Xingu, na Floresta Amazônica. O trabalho utiliza algoritmo genético para a tarefa de seleção de features e explorados parâmetros diferentes do algoritmo na busca por melhora na precisão da previsão. As features selecionadas pelo algoritmo são usadas como entrada de um modelo de Regressão Linear que realiza a previsão final. Uma análise estatística foi implementada pelos autores, a fim de comprovar que o modelo final pode prever o nível do rio com

alta precisão. Nesse contexto um coeficiente de determinação igual a 0,988 foi obtido. O Algoritmo Genético proposto mostrou-se portanto bem-sucedido na seleção das features mais relevantes para o problema.

3.1 Discussão dos Trabalhos

Analisando os trabalhos relacionados apresentados nesse Capítulo é possível notar que a utilização de sensores virtuais para lidar com lacunas e estimativas na coleta de dados tem se mostrado uma abordagem promissora e inovadora. Ao empregar técnicas de aprendizado de máquina, esses sensores são capazes de estimar e preencher os valores ausentes com base em correlações e padrões identificados nos dados disponíveis. Dessa forma, a integridade e a continuidade da análise são mantidas, mesmo diante da falta de informações em determinados momentos. Essa estratégia trás benefícios significativos para a detecção de anomalias e incremento da confiabilidade em sistemas e processos. Ao preencher as lacunas de dados, os sensores virtuais proporcionam uma visão mais completa do comportamento normal do sistema, permitindo uma identificação mais precisa e confiável de desvios ou comportamentos anômalos. Isso é especialmente relevante em contextos em que a ocorrência de anomalias pode ter consequências críticas, como na indústria, no setor energético ou na área da saúde. Uma das principais vantagens dos sensores virtuais é sua capacidade de adaptabilidade e aprendizado contínuo. Eles são capazes de ajustar suas estimativas e inferências com base em novos dados coletados ao longo do tempo, aprimorando ainda mais a precisão e a eficácia na detecção de anomalias. Além disso, a combinação de múltiplos sensores em um único modelo virtual contribui para uma análise mais holística e abrangente dos sistemas em estudo. Nos trabalhos discorridos, fica claro que a qualidade dos dados disponíveis desempenha um papel fundamental no desempenho dos sensores virtuais. Dados incompletos, ruidosos ou imprecisos podem afetar a acurácia das estimativas. Portanto, é necessário garantir a integridade e a confiabilidade dos dados utilizados no treinamento e na operação desses sensores virtuais. Em suma, a utilização de sensores virtuais na detecção de anomalias e em tratativas de gap de dados na coleta de informações representa uma abordagem avançada e inovadora. Com seu potencial de preencher as lacunas de dados, adaptabilidade e aprendizado contínuo, esses sensores têm o poder de melhorar a eficiência operacional, aumentar a segurança dos sistemas e fornecer insights valiosos em diversos setores.

Ao revisar os trabalhos relacionados, destaca-se que a utilização de sensores virtuais e as técnicas de machine learning, embora promissora, ainda não foram exploradas no contexto específico de sensores de instrumentação básica de barragens de rejeito. Nos estudos analisados, a ênfase recai sobre a aplicação de técnicas de aprendizado de máquina para lidar com lacunas e estimativas na coleta de dados, além de aumentar a confiabilidade de sistemas de previsão, proporcionando uma visão mais completa do comportamento dos

sistemas monitorados. No entanto, o diferencial do presente trabalho reside na adaptação dessas técnicas para o monitoramento específico de estruturas de contenção de rejeito de minério de ferro. A proposta busca preencher as lacunas de dados de forma adaptável, considerando as características particulares desse contexto, aprimorando assim a geração de sensores virtuais. Além disso, nosso estudo pretende abordar de maneira sistemática a questão da qualidade dos dados, garantindo a integridade e confiabilidade das informações utilizadas na implementação dos sensores virtuais. Ao fazê-lo, pretende-se contribuir significativamente para a eficácia operacional, segurança e tomada de decisões informadas nas operações relacionadas a estruturas de contenção de rejeito de minério de ferro, preenchendo uma lacuna identificada na literatura existente.

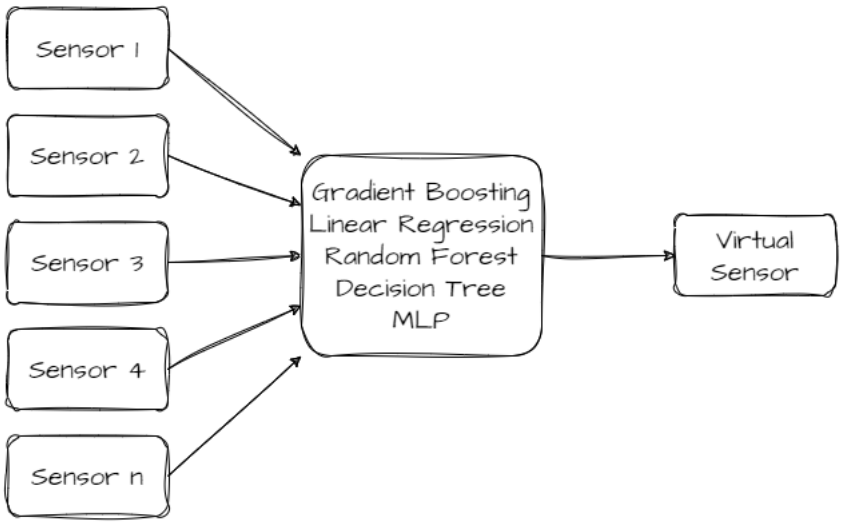
Capítulo 4

Método Proposto

4.1 Visão geral do trabalho

Com o objetivo de estimar a poropressão fornecida por piezômetros de corda vibrante presentes em barragens de rejeito, optou-se pela utilização de cinco modelos de machine learning: Randon Forest, MLP, Decision Tree, Linear Regression e Gradient Boosting. A última etapa consiste na avaliação e análise dos modelos propostos em termos de performance e acurácia. Abaixo temos o diagrama proposto para o método preliminar:

Figura 4.1: Visão geral do modelo proposto



Fonte: Os Autore

Em uma etapa final o modelo de maior robustez para o problema proposto será implementado dentro da camada PIMS, mais especificamente dentro do PI System.

4.2 Métricas de avaliação

Na fase de testes dos modelos de sensores virtuais, foram escolhidas métricas amplamente utilizadas para avaliar o desempenho de algoritmos: o Erro Médio Absoluto (Mean Absolute Error) (MAE), a Raiz do Erro Quadrático Médio (Root Mean Square Error) (RMSE) e o Coeficiente de determinação R^2. James (2013) fornece explicações detalhadas sobre as métricas de avaliação escolhidas, em um contexto de aprendizado de máquina e análise estatística. Abaixo, um resumo das informações fornecidas pelo livro sobre essas métricas:

- RMSE: métrica que avalia o quão bem um modelo de regressão ajusta os dados aos valores observados. Ele mede a raiz quadrada da média dos quadrados das diferenças entre os valores previstos pelo modelo e os valores reais. O RMSE atribui pesos maiores a erros maiores, o que significa que é mais sensível a erros grandes. Portanto, é adequado quando erros grandes são particularmente prejudiciais ou quando a distribuição dos erros é normal. A Equação (4.1) mostra como o indicador é calculado:

$$RMSE = \sqrt{\frac{1}{k}\sum_{m=1}^{k}(a_m - \hat{a}_m)^2} \tag{4.1}$$

  Onde "K"é o número de observações, a_m é o valor observado e $\hat{a}_m$ é o valor predito.

- MAE: métrica alternativa que avalia o desempenho de um modelo de regressão. Ele mede a média das diferenças absolutas entre os valores previstos e os valores reais. O MAE atribui pesos iguais a todos os erros, independentemente do tamanho, tornando-o menos sensível a valores atípicos do que o RMSE. Portanto, é útil quando se deseja uma métrica de erro mais robusta. O valor de MAE é calculado pela Equação (4.2):

$$MAE = \frac{1}{p}\sum_{q=1}^{p}|r_q - r| \tag{4.2}$$

  Onde, "p"representa o número de erros e $|r_q r|$ denota o erro absoluto.

- R^2: métrica que avalia a proporção da variabilidade na variável dependente que é explicada pelo modelo de regressão. Ele varia de 0 a 1, onde 0 indica que o modelo não explica nenhuma variabilidade e 1 indica que o modelo explica toda a variabilidade. O R^2 é uma métrica importante para entender o poder explicativo do

modelo. Quanto mais próximo de 1, melhor o modelo ajusta os dados. A Equação (4.3) representa sua proposta. ˜

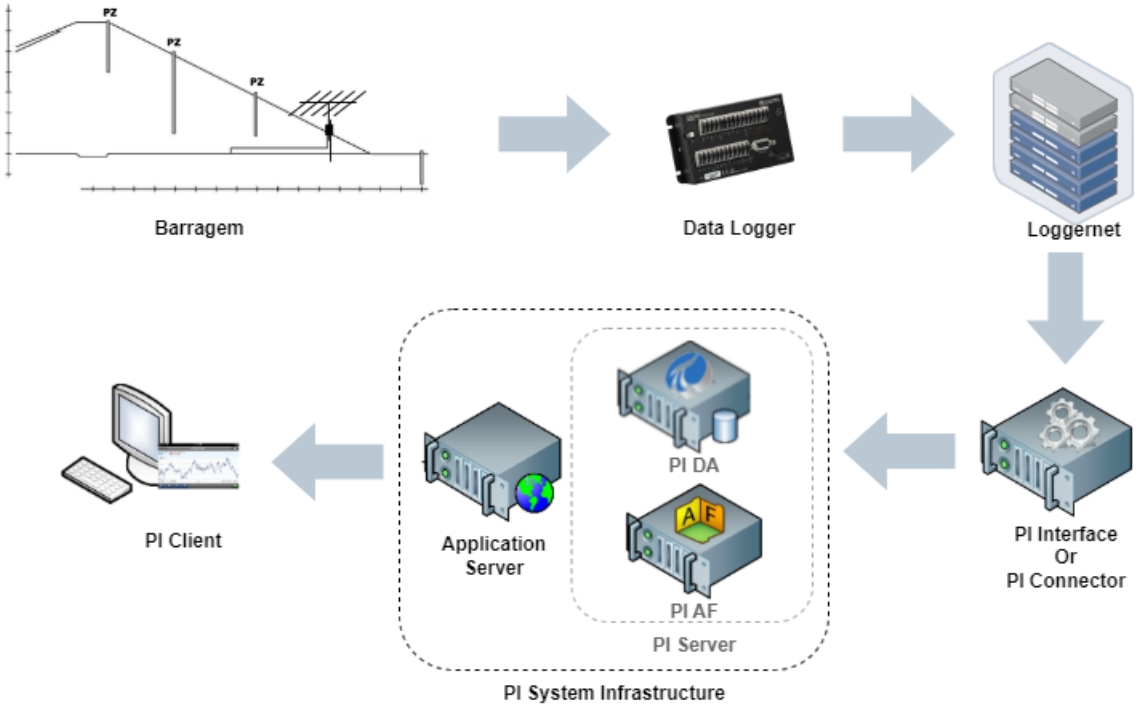
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(t_i - \hat{t}_i)^2}{\sum_{i=1}^{n}(t_i - \bar{t})^2} \tag{4.3}$$

Onde ti e o valor atual cumulativo de casos confirmados e $\hat{t}_l$ é o valor predito cumulativo de casos confirmados.

4.3 Coleta e Fluxo de dados

O fluxo de dados geral proposto para este trabalho está demonstrado na Figura 4.2, onde temos no início do fluxo a representação de uma barragem de rejeitos contendo múltiplos piezômetros distribuídos na estrutrura. Os sinais elétricos coletados pelos piezômetros são traduzidos e armazenados pelo Data Logger, posteriormente são transmitidos via antena rádio e disponibilizados em uma máquina central responsável pelo gerenciamento da informação através do software Loggernet. A PI Interface/Connector for Loggernet é responsável por colher essa informação contida no Loggernet e persistí-la em forma de Tag dentro do PI Data Archive (PI DA). Por sua vez o PI Asset Framework (PI AF) pode gerar uma camada de contextualização da informação em forma de árvore e atributos, desse modo os dados podem ser visualizados de forma melhor pelos usuários de interesse, representados pelas ferramentas cliente oferecidas pelo PI System (PI Client). Por fim temos a representação de um servidor de aplicação dentro da infraestrutura do PI System, onde uma aplicação Python integrada ao PI AF SDK pode ser abrigada para implementação de modelos de aprendizado de máquina.

Figura 4.2: Fluxo de dados



Fonte: Os Autores

4.3.1 Data Set

Nesta etapa do projeto o dataset utilizado refere-se a coleta de dados de doze piezômetros elétricos de corda vibrante, por se tratar de um trabalho investigativo de métodos de aprendizado de máquina, foram utilizados dados de uma barragem de alto nível de segurança. O conjunto de dados em questão possui 3901 amostras obtidas no período entre 17-jan-2023 11:00:00 e 02-jul-2023 17:00:00, onde cada sensor possui taxa de coleta igual a uma hora.

Capítulo 5

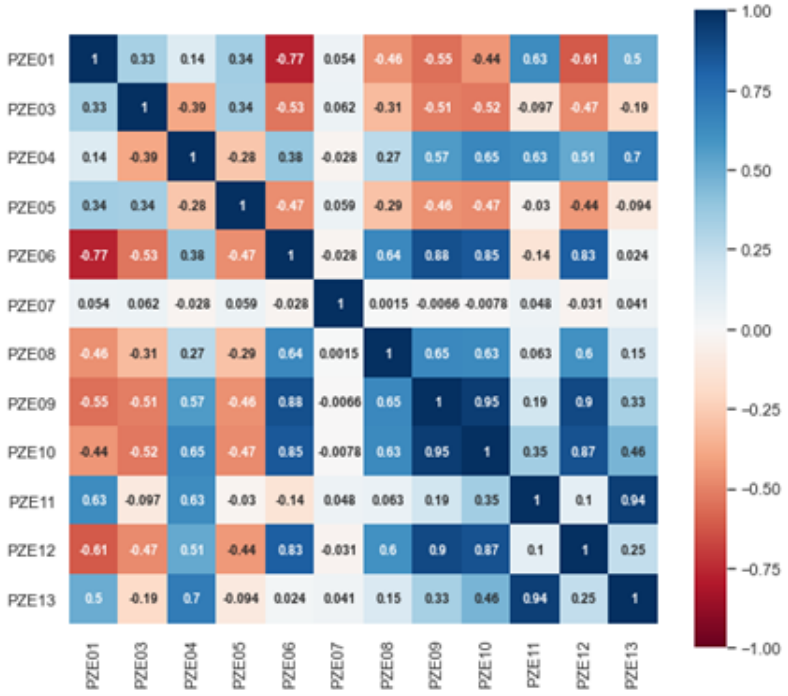
Investigação e Resultados

5.1 Mapa de Correlações

Com o objetivo de estabelecer a correlação entre os piezômetros, foram construídos dois mapas de correlação, o primeiro usando o método de Pearson e o segundo utilizando o método de Spearman. Um mapa de correlações é uma representação visual das correlações entre variáveis, representando uma medida estatística que descreve a relação entre duas variáveis, indicando se elas estão associadas e em que grau.

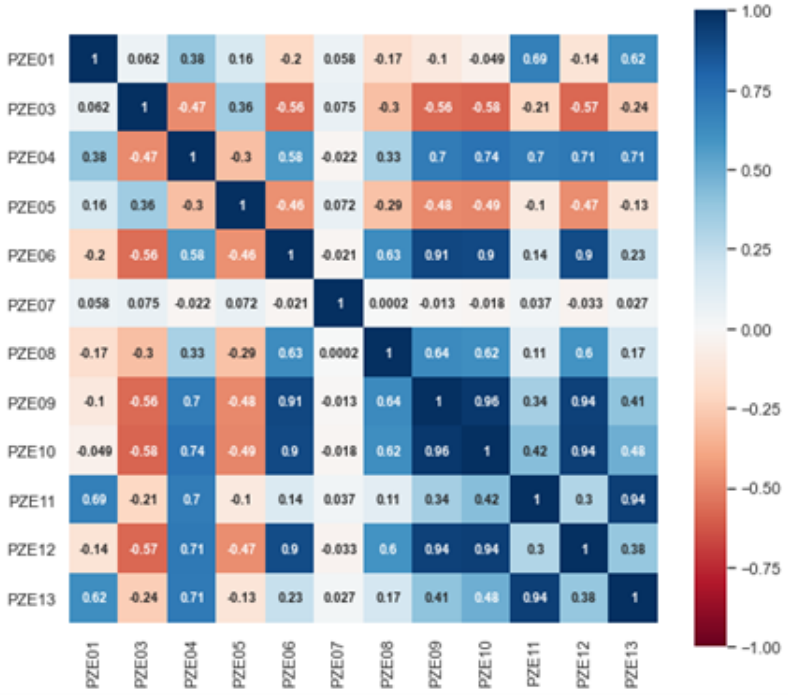
Os mapas utilizados neste trabalho correlacionam os piezômetros PZE01, PZE03, PZE04, PZE05, PZE06, PZE07, PZE08, PZE09, PZE10 PZE11, PZE12 e PZE13, portanto a diagonal principal dos mapas é dada por 1, situação na qual um piezômetro está correlacionado com ele mesmo.

Figura 5.1: Mapa de correlações entres os PZs usando método de Pearson



Fonte: Os Autores

Figura 5.2: Mapa de correlações entres os PZs usando método de Spearman

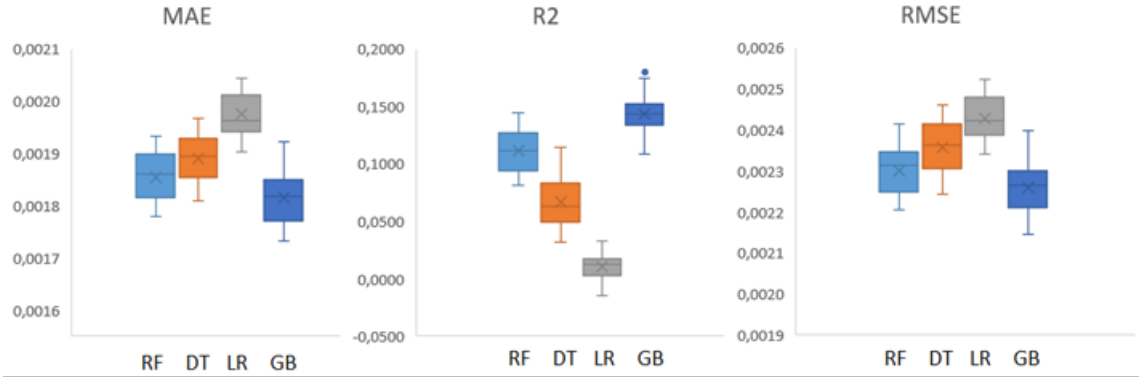


Fonte: Os Autores

5.2 Investigação sobre modelo para estimativa do PZE com menor correlação entre os vizinhos

Como observado nos mapas de correlação das Figuras 5.1 e 5.2, o piezômetro que apresentou a menor correlação entre seus vizinhos foi o PZE07. Foram avaliadas cinco técnicas de aprendizado de máquina para estimativa do valor do PZ07. As técnicas utilizadas foram Redes Neurais, Árvores de Decisão, Regressão Linear, Random Forest e Gradient Boosting. O processo de treino e teste dos modelos foi feito 30 vezes, considerando 80% dos dados para treino e 20% para teste. A entrada dos modelos considera todos os demais piezômetros disponíveis PZE01, PZE03, PZE04, PZE05, PZE06, PZE08, PZE09, PZE10, PZE11, PZE12, PZE13, sendo a saída do modelo o valor esperado para PZ07. Os dados foram obtidos de 12 piezômetros, na barragem de Forquilha V, com frequência horária, entre os dias de 17-jan-2023 e 02-jul-2023. Utilizou-se inicialmente os seguintes parâmetros nas técnicas de aprendizado de máquina: Random Forest (20 árvores, profundidade máxima = 5). Decision Tree (max_depth=5). Neural Network (MLPRegressor, hidden_layer_sizes=(8,8), activation='tanh', solver='lbfgs', learning_rate_init=0.001, max_iter=5000, momentum=0.9). Gradient Boosting Regressor (n_estimators = 20, max_depth = 5).

Figura 5.3: Resultado de 30 rodadas de treino e teste. Métricas de avaliação MAE, R2 e RMSE.



Fonte: Os Autores

A Figura 5.3 apresenta o resultado de 30 rodadas de treino e teste, apresentando as métricas de avaliação MAE, R^2 e RMSE para os modelos Random Forest, Decision Tree, Linear Regression e Gradient Boosting, o box plot referente ao Multilayer Perceptron foi removido, pois aprensetou resultados muito inferiores aos demais, comprometendo a visualização do gráfico. Podemos ver pelo R^2 (quanto mais próximo de 1 melhor) que o coeficiente de determinação para todos os modelos é muito baixo, inferior a 0,2. O resultado do R^2 pode ser melhor entendido com os gráficos de linha das Figuras 5.4 e 5.5 que apresentam o valor esperado e o valor obtido pelos modelos Random Forest e Linear Regression. Podemos ver que praticamente não existe aprendizado. O que ocorre é que o

modelo estima basicamente a média geral da série.

Figura 5.4: Valores esperados e obtidos pelos modelos usando Random Forest. Pelo R2 medido, podemos dizer que não houve aprendizado.

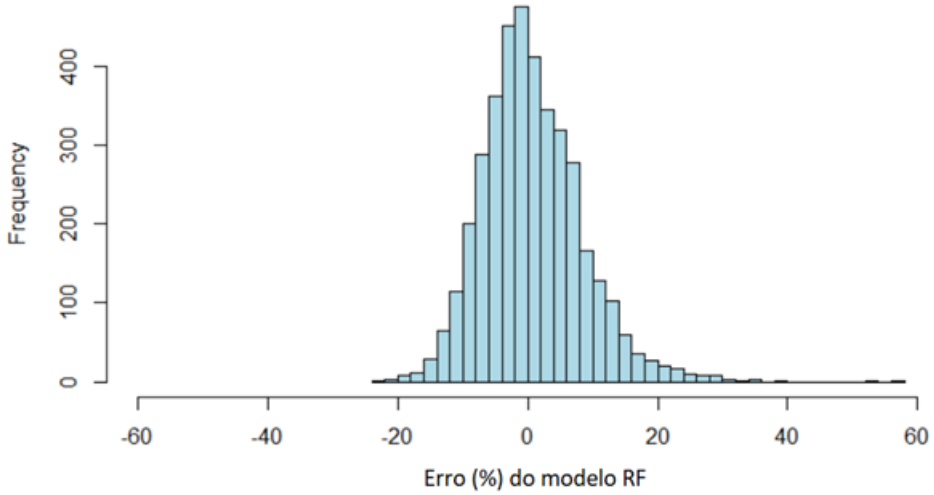
Fonte: Os Autores

Figura 5.5: Valores esperados e obtidos pelos modelos usando Regressão Linear. Pelo R2 medido, podemos dizer que não houve aprendizado.

Fonte: Os Autores

Podemos notar que praticamente não existe aprendizado, uma vez que o modelo estimou basicamente a média geral das séries. A Figura 5.6 apresenta o erro, medido em %, da diferença entre o valor obtido pelo modelo RF em relação aos dados esperados. Erro varia entre -23% e 56%.

Figura 5.6: Erro (%) do valor obtido pelo modelo de RF em relação aos dados esperados.



Fonte: Os Autores

Umas das hipóteses levantadas é que, como a correlação entre vizinhos é muito fraca, não existem informações suficientes para gerar aprendizado com os dados utilizados.

5.3 Investigação sobre modelo para estimativa do PZE com maior correlação entre os vizinhos

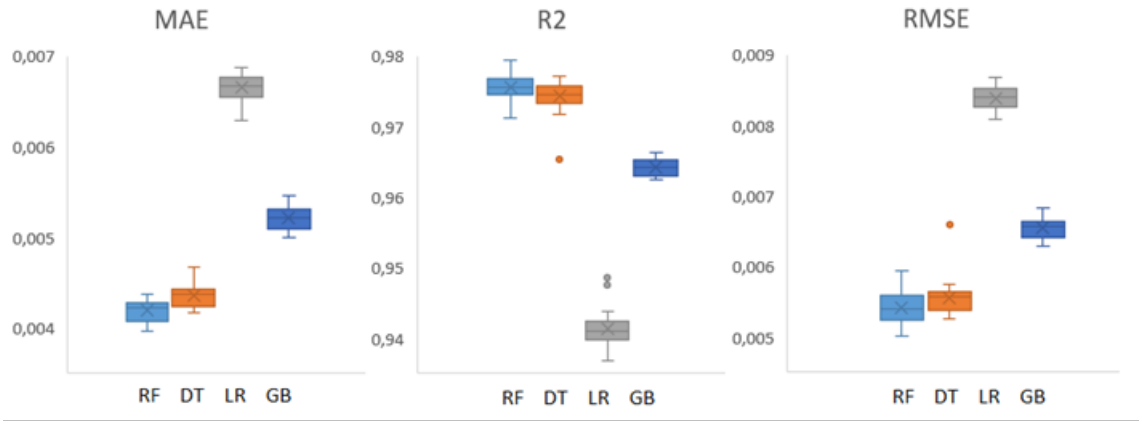
Realindo o mesmo protocolo descrito na seção 5.2, porém tomando como objetivo estimar o valor do PZ que apresentou a maior correlação entre os vizinhos, de acordo com o mapa de correlações PZE06. Foram avaliadas cinco técnicas de aprendizado de máquina para estimativa do valor do PZ06. As técnicas utilizadas foram Redes Neurais, Árvores de Decisão, Regressão Linear, Random Forest e Gradient Boosting. O processo de treino e teste dos modelos também foi feito 30 vezes, considerando 80% dos dados para treino e 20% para teste. A entrada dos modelos considera todos os demais piezômetros disponíveis PZE01, PZE03, PZE04, PZE05, PZE07, PZE08, PZE09, PZE10, PZE11, PZE12, PZE13, sendo a saída do modelo o valor esperado para PZE06. Utilizamos inicialmente os seguintes parâmetros nas técnicas de aprendizado de máquina:

Tabela 5.1: Parâmetros utilizados nos métodos de aprendizado de máquina propostos.

	N° de Árv	Prof.Máx	N° Estm	Mtdo	Tx.AprendIni	Cam.Ocultas	Fun.Ativ	Solver	Moment
Random Forest	20	5	-	-	-	-	-	-	-
Neural Network	-	-	-	MLP	0,001	8,8	tanh	lbfgs	0,9
Decision Tree	-	5	-	-	-	-	-	-	-
Gradient Boosting	-	5	20	-	-	-	-	-	-

A Figura 5.7 apresenta o resultado de 30 rodadas de treino e teste, apresentando as métricas de avaliação MAE, R^2 e RMSE para os modelos Random Forest, Decision Tree, Linear Regression e Gradient Boosting, o box plot referente ao Multilayer Perceptron

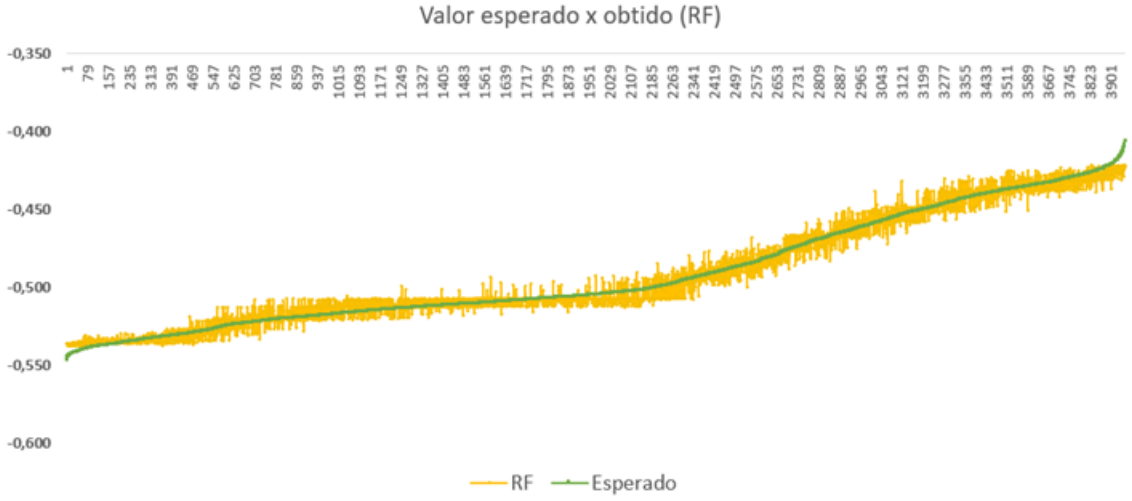
Figura 5.7: Resultado de 30 rodadas de treino e teste. Métricas de avaliação MAE, R2 e RMSE.



Fonte: Os Autores

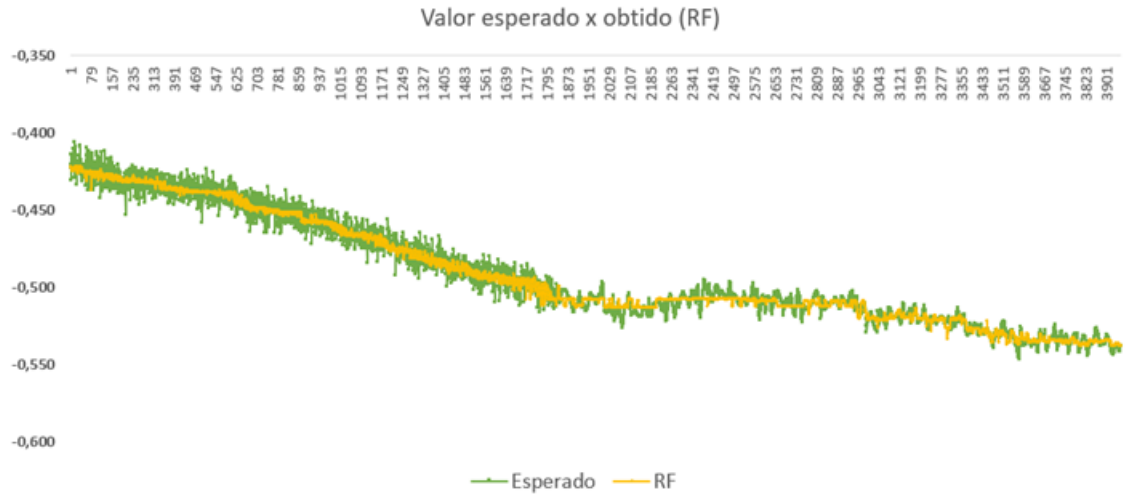
foi removido, pois aprensetou resultados muito inferiores aos demais, comprometendo a visualização do gráfico. Podemos ver pelo R^2 (quanto mais próximo de 1 melhor) que o coeficiente de determinação para todos os modelos foi alto, superior a 0,96 tanto para o RF, como DT e . Apenas para LR ficou abaixo de 0.96. O resultado das métricas pode ser melhor entendido com o gráfico de linha na Figura 5.8 que apresenta o valor esperado e o valor obtido pelos modelos. Podemos ver que os resultados obtidos pelo RF seguem a linha dos dados esperados. (ajustar) Dados não foram normalizados, assim, a comparação entre PZ06 e PZ07 neste caso só pode ser feito por meio do R2 e dos resultados em percentual. A comparação não pode ser feita por meio do MAE ou RMSE devido a estes serem referentes a escala de cada PZ.

Figura 5.8: Valores esperados e obtidos pelo modelo RF, ordenando o dataset pelo valor dos resultados esperados. Note que o modelo segue a linha, entretanto, próximo aos limites inferior e superior não é apresentada a mesma assertividade do modelo.



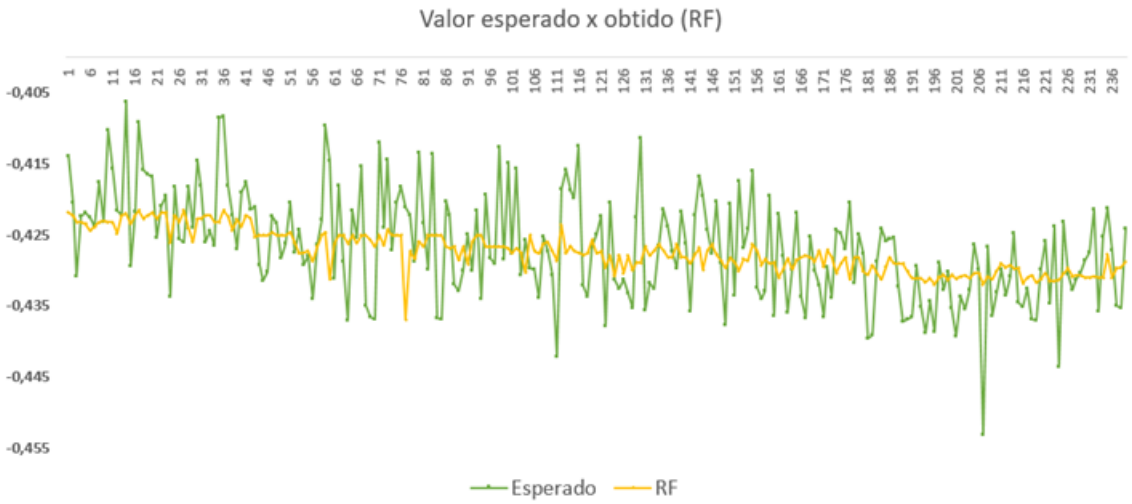
Fonte: Os Autores

Figura 5.9: Valores esperados e obtidos pelo modelo RF usando a sequência natural do dataset. Esta visualização nos dá a impressão de que os dados do PZs tem certo ruído, e que o modelo faz algo como uma média móvel.



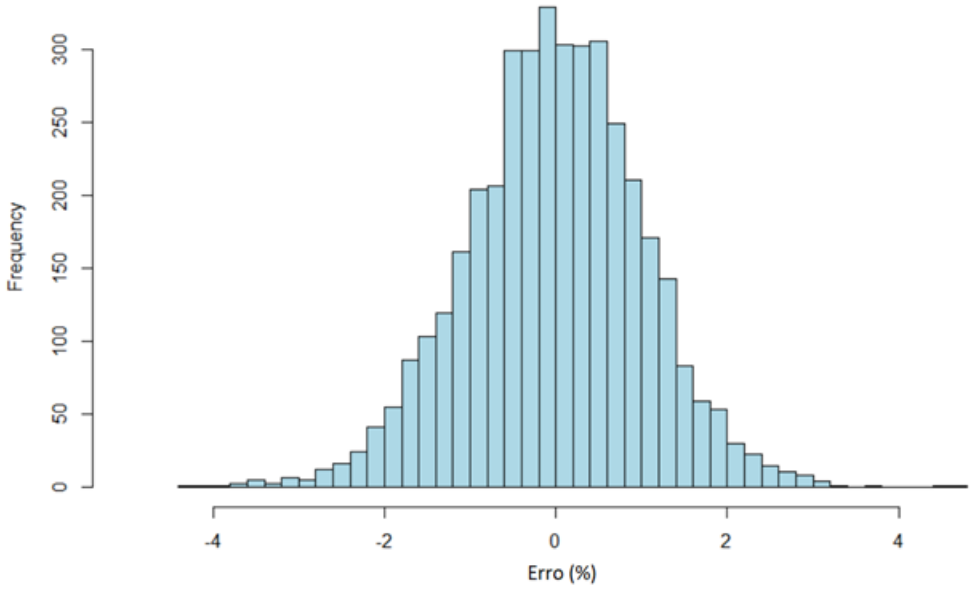
Fonte: Os Autores

Figura 5.10: Valores esperados e obtidos pelo modelo RF usando a sequência natural do dataset. Similar a Figura 5.9, porem apresentando zoom nas primeiras 240 horas.



Fonte: Os Autores

Figura 5.11: Erro (%) do valor obtido pelo modelo de RF em relação aos dados esperados.



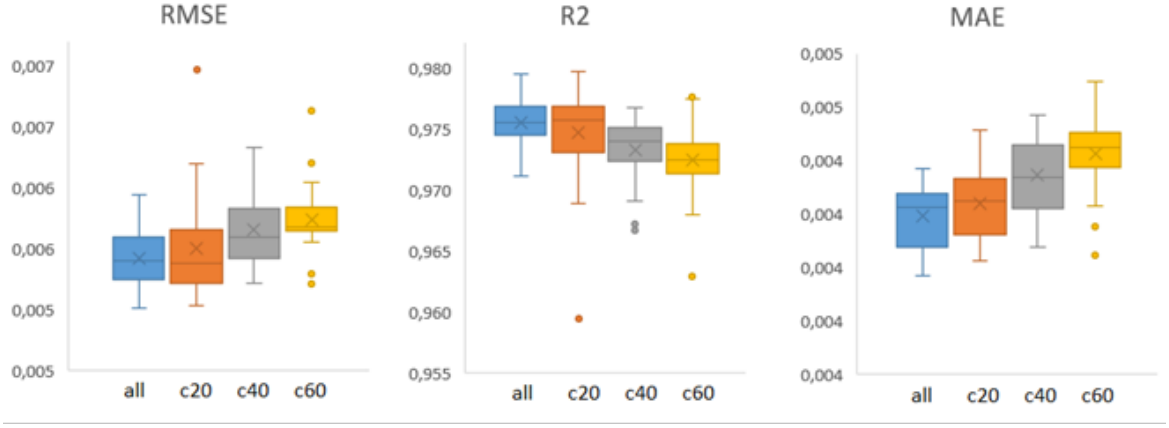
Fonte: Os Autores

O erro variou de $-4,26\%$ e $4,69\%$

5.4 Investigação sobre grau de correlação dos vizinhos

Buscamos avaliar como diferentes graus de correlação afetam o aprendizado. Conduzimos um experimento considerando todos os vizinhos e removendo os com correlação inferior a 20, inferior a 40 e inferior a 60. Nesta investigação, usamos apenas a técnicas de aprendizado que apresentou os melhores resultados na seção anterior (Random Forest, 20 árvores, 5 níveis). Usamos como avaliação o PZ06. A Figura 5.5 apresenta o RMSE, R2 e MAE estimando PZ06 considerando todos os vizinhos (all) e remoção dos vizinhos com correlação inferior a 0.20 (c20), 0,40 (c40), e 0,60 (c60).

Figura 5.12: RMSE, R2 e MAE estimando PZ06 considerando todos os vizinhos (all) e remoção dos vizinhos com correlação inferior a 0.20 (c20), 0,40 (c40), e 0,60 (c60).



Fonte: Os Autores

Buscando uma avaliação estatística, realizamos (1) teste de normalidade dos conjuntos e (2) teste de similaridade entre os conjuntos.

Após observar em parte dos conjuntos uma distribuição não aderente à normal com o teste de Shapiro-Wilk, optou-se pela utilização de um teste não paramétrico de similiridade. A tabela 5.3 apresenta o teste de similaridade de Wilcoxon aplicado ao conjunto de dados obtido.

Nota-se que MAE, RMSE e R2 são similares para os conjuntos de correlação C20 e All, indicando que a exclusão de piezômetros com correção inferiores a 20 pode ser utilizada para efeitos de redução do conjunto de dados avaliados, consequentemente redução do custo computacional atribuído a execução do modelo avaliado.

Tabela 5.2: Avaliação de Correlação: Teste de normalidade dos conjuntos, usando o Shapiro-Wilk normality test. Valores em negrito são os com p-valor > 0,05 (considerados como aderentes a distribuições normais).

Conjunto	**p-Valor**
RMSEc20	0,000
RMSEc40	0,027
RMSEc60	0,004
RMSEall	**0,840**
MAEc20	**0,319**
MAEc40	**0,173**
MAEc60	0,015
MAEall	0,030
R2c20	0,000
R2c40	0,008
R2c60	0,005
R2all	**0,753**

Fonte: Os Autores

Tabela 5.3: Avaliação de Correlação: Teste de similaridade dos conjuntos, usando o Wilcoxon rank sum test with continuity correction. Valores em negrito são os com p-valor > 0,05 (considerados como similares).

Wilcoxon rank sum exact test	**p-Valor**
RMSEc20,RMSEall	**0,687**
RMSEc40,RMSEall	0,002
RMSEc60,RMSEall	0,000
MAEc20,MAEall	**0,202**
MAEc40,MAEall	0,000
MAEc60,MAEall	0,000
R2c20,R2all	**0,820**
R2c40,R2all	0,002
R2c60,R2all	0,000

Fonte: Os Autores

5.5 Investigação sobre tamanho de janelas (time lag) e parâmetros do Random Forest

Com o objetivo de melhorar os resultados da estimativa, foi realizada uma segunda avaliação. Utilizamos a melhor arquitetura encontrada na etapa anterior e avaliamos diferentes tamanhos para a janela de entrada, ou seja, diferentes quantidades de dados anteriores para prever estimar valores atuais. A tabela5.4 mostra a estruturação de janelas avaliadas:

A Figura 5.13 apresenta um diagrama esquemático com valores genéricos de entradas e saídas a serem usados no modelo de estimativa. A Figura mostra o modelo utilizado,

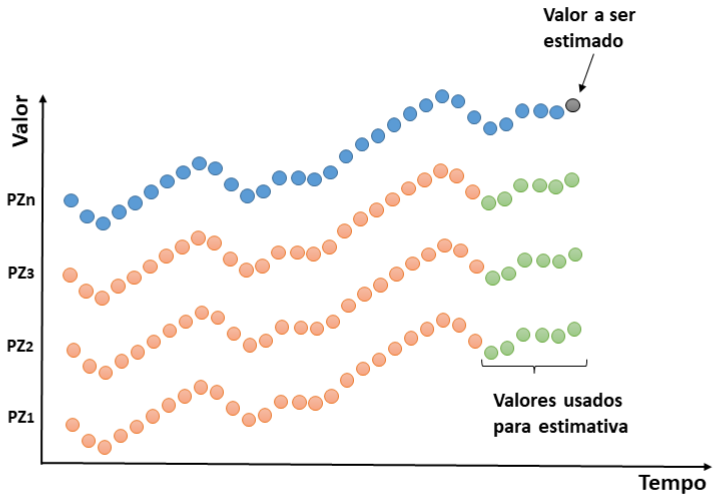
Tabela 5.4: Estrutura de avaliação de janelas temporais dos dados de entrada. Relacionado a entrada, t_0 significa o tempo atual. t_2 significa 3 horas antes da amostra estimada no tempo atual, t_5 significa 6 horas antes da amostra estimada no tempo atual

Referência	**Timestamp entrada**	**Timestamp saída**
w1	t_0	t_0
w3	t_0, $t-1$, $t-2$	t_0
w6	t_0, $t-1$, $t-2$, $t-3$, $t-4$, $t-5$	t_0

Fonte: Os Autores

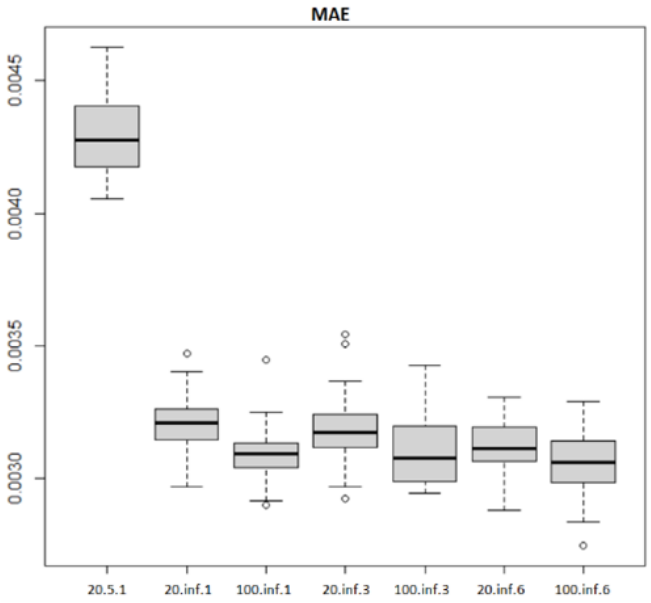
embora não represente dados reais coletados pelo sistema. O objetivo é representar graficamente a maneira pela qual os valores são empregados. Nesta Figura, representamos a janela w6, que é um vetor composto pelo tempo atual (t0), 1, 2, 3, 4 e 5 horas anteriores ao evento a ser estimado.

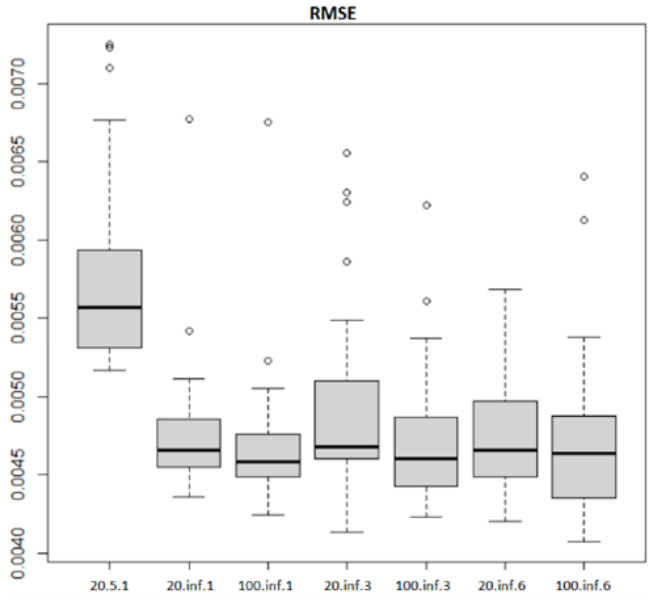
Figura 5.13: Diagrama de representação genérico dos valores de entradas e saída usados no modelo. No qual temos em cinza o valor estimado no instante t0 para o Piezômetro PZn, em relação às janelas temporais em t-5 (W6), sinalizada pelos valores em verde.

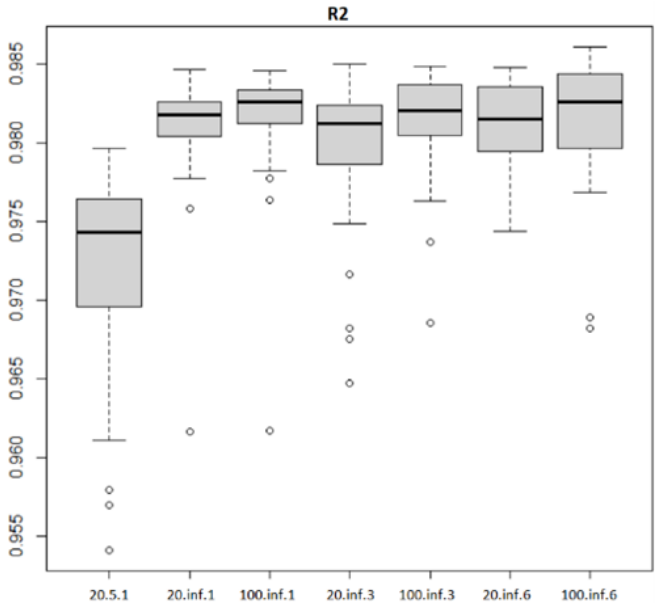


Fonte: Os Autores

Figura 5.14: RMSE, R2 e MAE estimando PZ06 considerando diferentes números de árvores, diferentes janelas e profundidade das árvores. Conjuntos são descritos como [número de árvores-profundidade-janela].







Fonte: Os Autores

Notou-se que não limitar a profundidade das árvores trouxe melhora no desempenho dos modelos, reduzindo o MAE e o RMSE e aumentando o R2 do Random Forest.

Tabela 5.5: Avaliação de Janelas Temporais e Parâmetros do Random Forest: Teste de normalidade dos conjuntos, usando o Shapiro-Wilk normality test. Valores em negrito são os com p-valor > 0,05 (considerados como aderentes a distribuições normais).

Conjunto	**p-Valor**
RMSE-100-inf-1	0,0000
RMSE-100-inf-3	0,0004
RMSE-100-inf-6	0,0002
RMSE-20-5-1	0,0002
RMSE-20-inf-1	0,0000
RMSE-20-inf-3	0,0001
RMSE-20-inf-6	**0,1395**

Fonte: Os Autores

Tabela 5.6: Avaliação de janelas temporais e parâmetros do Random Forest: Teste de similaridade dos conjuntos, usando o Wilcoxon rank sum test with continuity correction. Valores em negrito são os com p-valor > 0,05 (considerados como similares).

Conjunto	**p-Valor**
(RMSE.20.inf.1,RMSE.20.inf.3)	**0,3516**
(RMSE.20.inf.1,RMSE.20.inf.6)	**0,9411**
(RMSE.100.inf.1,RMSE.100.inf.3)	**0,9882**
(RMSE.100.inf.1,RMSE.100.inf.6)	**0,8708**
(RMSE.20.inf.1,RMSE.100.inf.6)	**0,2928**

Fonte: Os Autores

Notou-se pelo teste de similaridade que os modelos não apresentaram melhora por meio da implementação de memória temporal no aprendizado dos métodos, umas vez que pelo teste de similaridade de Wilcoxon houve similarida da métrica RSME entre testes considerando janelas de 1, 3 e 6 horas. Houve similaridade também entre testes de 20 e 100 árvores para o modelo.

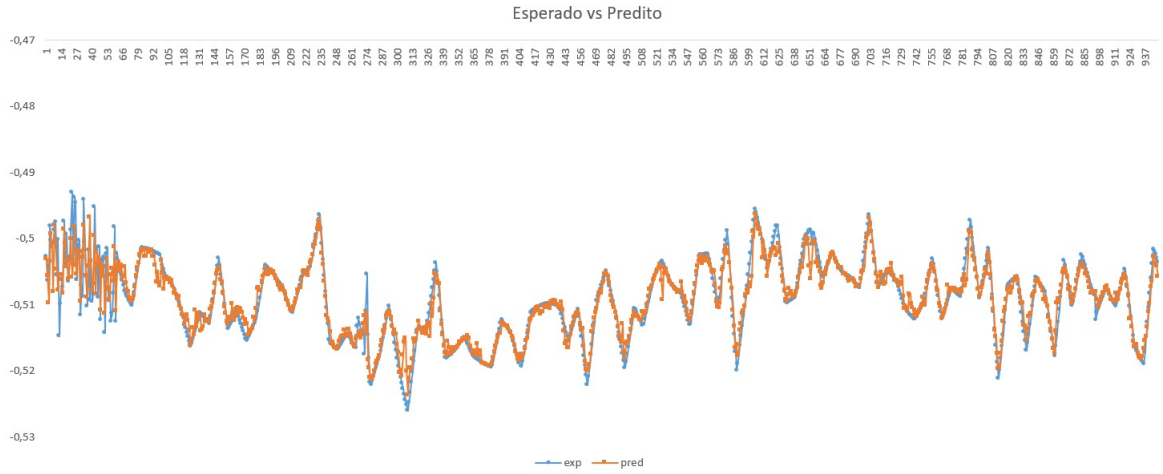
Uma das hipóteses é que os dados possuem padrão de longo prazo, que não podem ser efetivamente capturados por uma janela de tempo limitada, portanto a adição de janelas temporais não forneceu informação suficiente para melhorar o modelo.

5.6 Modelo Selecionado

Considerando os testes realizados nas seções anteriores, por seleção o modelo final Random Forest (20 árvores, profundidade máxima = inf), considerando os piezômetros com

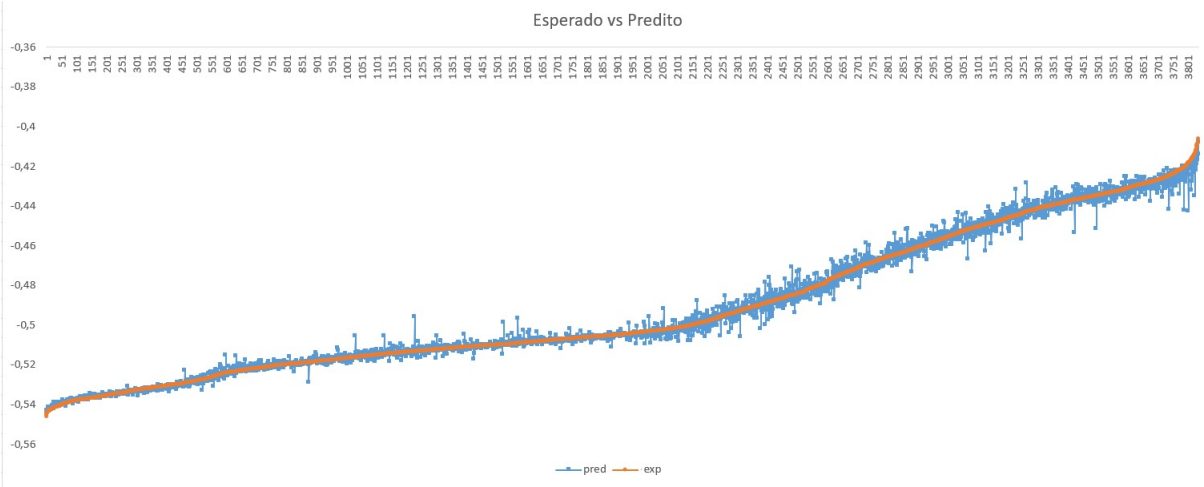
correlação superior a 20, e também a ausência de janelas temporais foram obtidos os resultados expressos pelos gráficos 5.15 e 5.16

Figura 5.15: Valores Esperados X Valores Obtidos (Primeiras Amostras)



Fonte: Os Autores

Figura 5.16: Valores Esperados X Valores Obtidos



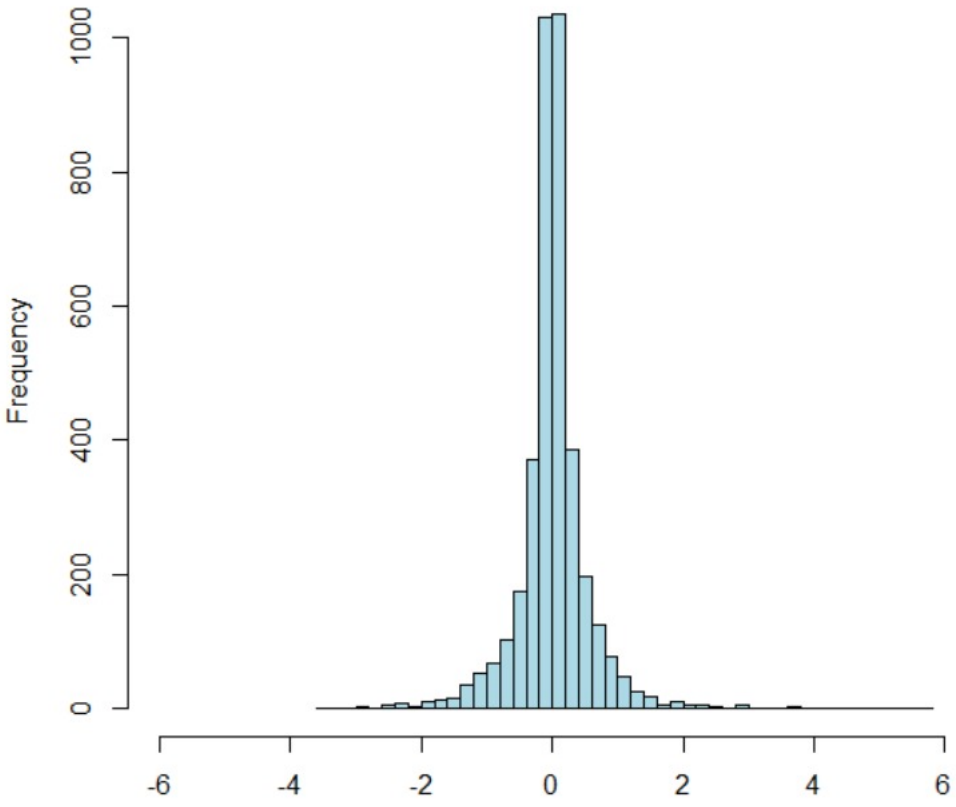
Fonte: Os Autores

Tabela 5.7: Métricas de performance: Melhor Modelo - Random Forest (20 árvores, profundidade máxima = inf), considerando os piezômetros com correlação superior a 20, e também a ausência de janelas temporais

	MAE	**RMSE**	R^2
Média de 30 execuções	0,003151	0,004715	0,981171
Melhor	0,002930	0,004231	0,985481

Fonte: Os Autores

Figura 5.17: Erro (%) do valor obtido pelo modelo final em relação aos dados esperados.



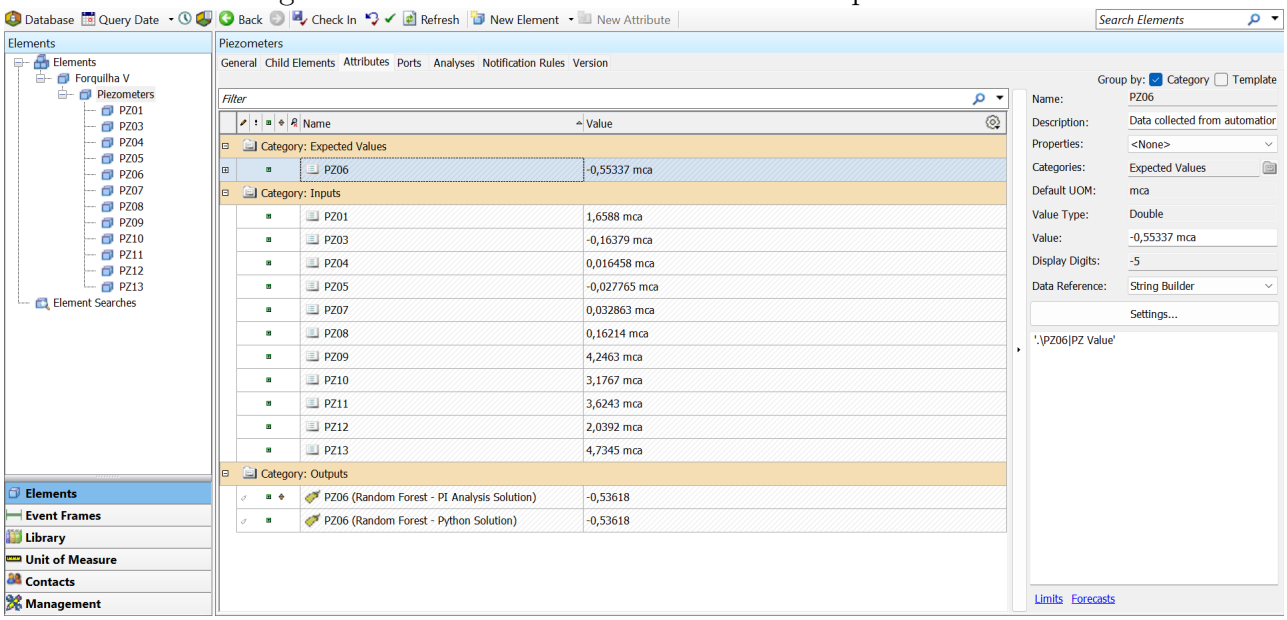
Fonte: Os Autores

Capítulo 6

Implementação PIMS

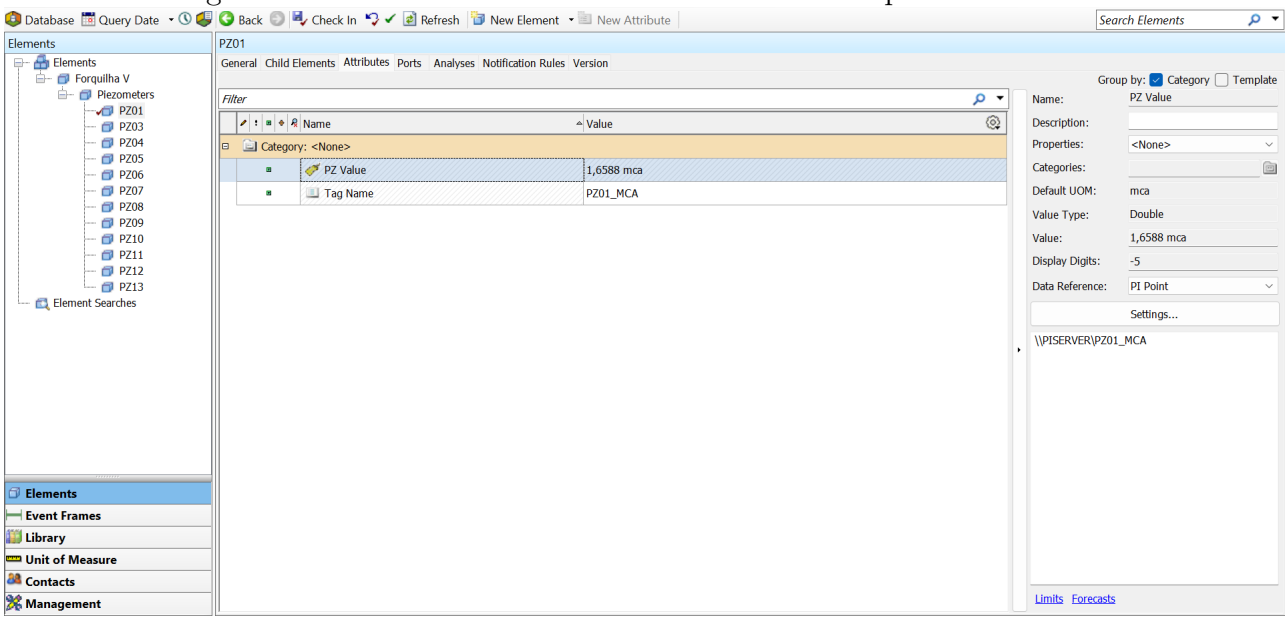
Para implementação dos resultados obtidos na camada PIMS, inicialmente gerou-se uma estruturação dos piezômetros no PI AF, de modo a criar uma camada de contextualização e visualização dos dados, conforme demonstrado nas Figuras 6.1 e 6.2. Os piezômetros foram incluídos na forma de atributos e categorizados de acordo com sua função no modelo. Temos portanto as entradas avaliadas, categorizadas como *Inputs*, o piezômetros *PZ06* contendos os valores esperados, representado portanto na categoria de *Expected Values* e por fim os valores obtidos representados na categoria de *Outputs* pelos atributos do tipo PI Point *PZ06 (Random Forest - PI Analysis Solution) e PZ06 (Random Forest - Python Solution).*

Figura 6.1: Árvore PI AF: Piezômetros Forquilha V.



Fonte: Os Autores

Figura 6.2: Árvore PI AF: Atributos vinculados ao piezômetro.



Fonte: Os Autores

Foram selecionadas para este trabalho duas formas de implementação do modelo final dentro da camada PIMS:

1. Integração via aplicação customizadas em PI AF SDK: nesse método é possível integrar modelos desenvolvidos em Python com o PI System, dessa forma pode-se armazenar os valores estimados dentro de um ponto de controle qualquer;

2. Implementação da árvore gerada na execução do Random Forest dentro do PI Analysis: nesse método de integração é possível transcrever as regras expressas pela árvore gerada na execução do Random Foreste dentro do PI Analysis.

6.1 Implementação via PI Analysis

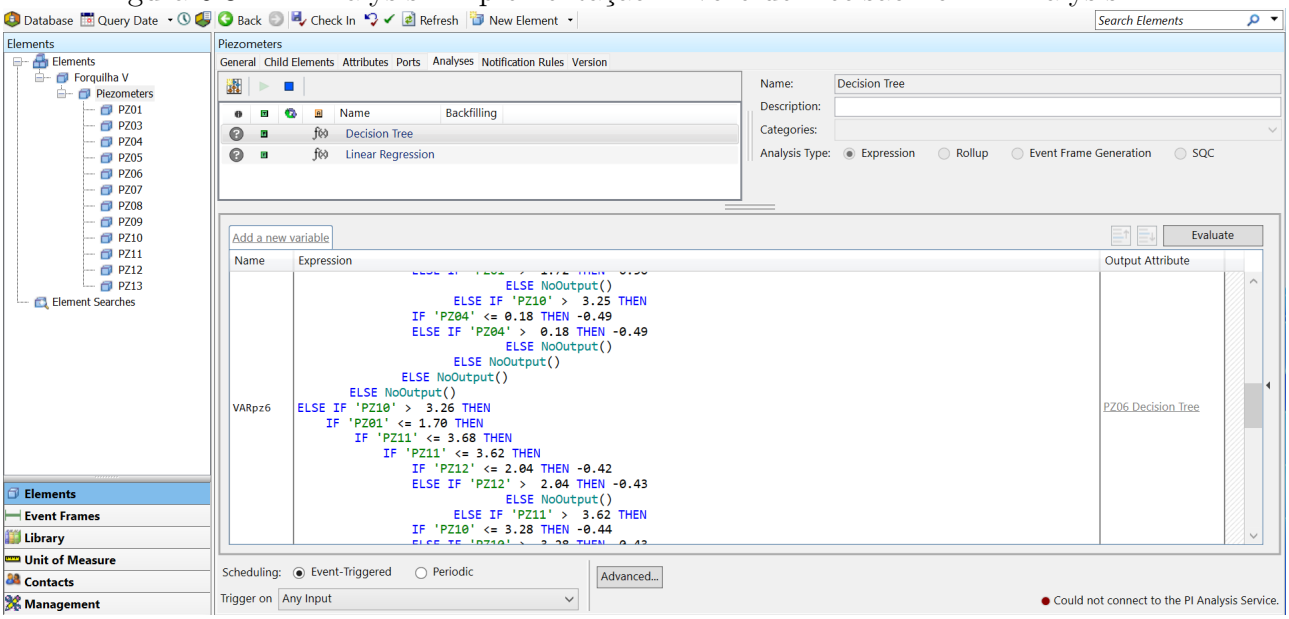
Para implementação via PI Analysis a árvore gerada pela execução do Random Forest foi simplificada e ajustada para atender a sintaxe do PI Analysis, conforme trechos de código abaixo:

Algoritmo 1: Random Forest Tree

Data: Entrada: PZ01, PZ03, PZ04, PZ05, PZ08, PZ09, PZ10, PZ12
Result: Saída: PZ06

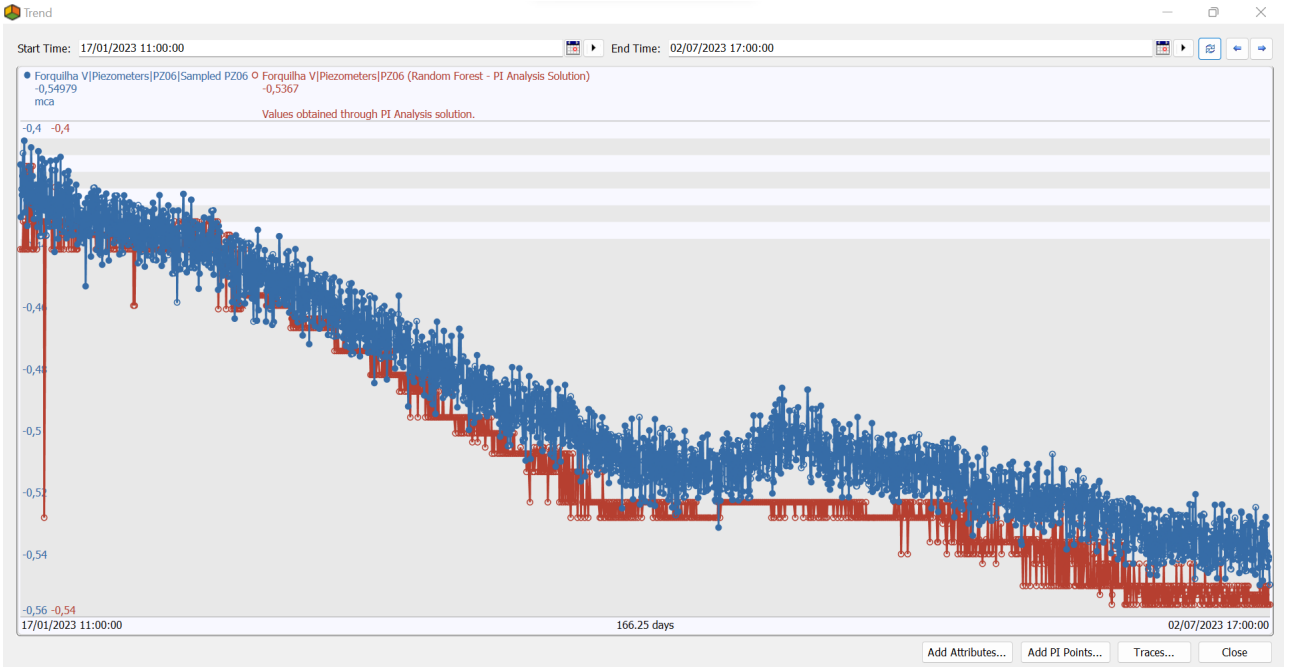
```
1  if 'PZ10' < 3.2629801035 then
2      if 'PZ10' < 3.2160348895 then
3          if 'PZ10' < 3.2070509195 then
4              if 'PZ10' < 3.198621631 then
5                  if 'PZ12' < 2.0150520805 then
6                      -0.5367022171
7                  else
8                      -0.5339137023
9              else
10                 if 'PZ09' < 4.2737691405 then
11                     -0.5314108957
12                 else
13                     -0.5251348997
14         else
15             if 'PZ08' < 0.16881042 then
16                 if 'PZ05' < -0.0327175445 then
17                     -0.5182222633
18                 else
19                     -0.5224070774
20             else
21                 if 'PZ05' < -0.025089965 then
22                     -0.5190208122
23                 else
24                     -0.5137658927
25     else
26         ...
27 else
28     ...
```

Figura 6.3: PI Analysis: Implementação Árvore de Decisão no PI Analysis



Fonte: Os Autores

Figura 6.4: Valores Esperados X Valores Obtidos - PZE06 PI Analysis Solution



Fonte: Os Autores

6.2 Implementação via Python

Para implementação via Python o código abaixo foi desenvolvido e integrado com o PI System, através da biblioteca PI AF SDK. Os valores estimados pelo modelo são ao final do código inputados no Tag de saída, representado pelo atributo do PI AF *"PZE06 (Random Forest - Python Solution)"* da Figura6.1.

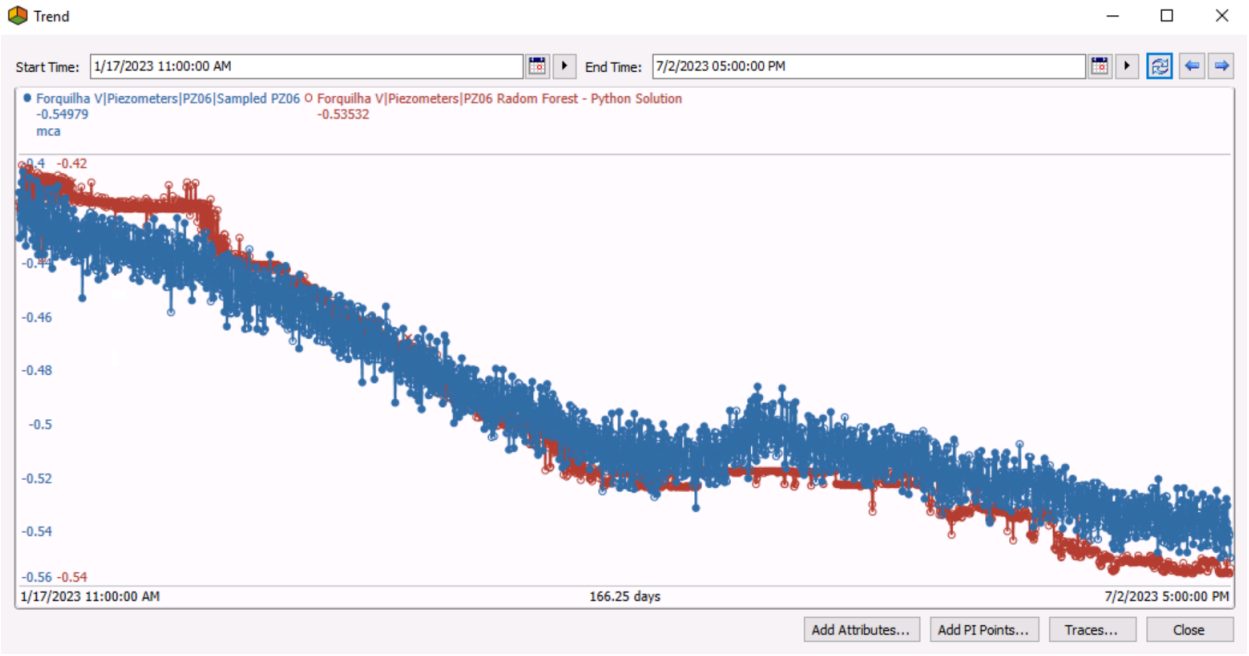
Algoritmo 2: Random Forest Algorithm

Data: Input Data: PZ01, PZ03, PZ04, PZ05, PZ08, PZ09, PZ010, PZ012
Result: Output: PZ06

```
// Getting Data
X ← [[PZ01, PZ03, PZ04, PZ05, PZ08, PZ09, PZ10, PZ12]].values;
y ← [PZE06].values;
// Splitting the dataset for training and testing
X_train, X_test, y_train, y_test ← train_test_split(X, y, test_size=.2);
// Fitting Random Forest Regressor Model
model_rf ← RandomForestRegressor(n_estimators=20);
model_rf.fit(X_train, y_train);
estimated_rf ← model_rf.predict(X_test);
expected ← y_test;
// Evaluating on remaining data
rmse_rf ← mean_squared_error(expected, estimated_rf, squared=False);
r2_rf ← r2_score(expected, estimated_rf);
mae_rf ← mean_absolute_error(expected, estimated_rf);
// Simulate use of models on entire dataset
estimatedFull_rf ← model_rf.predict(X);
expectedFull ← y;
// find PI Point by name
writept ← PIPoint.FindPIPoint(piServer,
  "PZ06(RandomForest − PythonSolution)");
// Write data to PI Point
writept.UpdateValues(estimatedFull_rf, AFUpdateOption.Replace,
  AFBufferOption.BufferIfPossible);
```

Na Figura 6.5 podemos ver os valores estimados para os dados de PZ06 onde o modelo implementado em Python escreve diretamente os resultados via PI AF SDK no Tag PIMS, representado pela curva em vermelho.

Figura 6.5: Valores Esperados X Valores Obtidos - PZE06 Python Solution

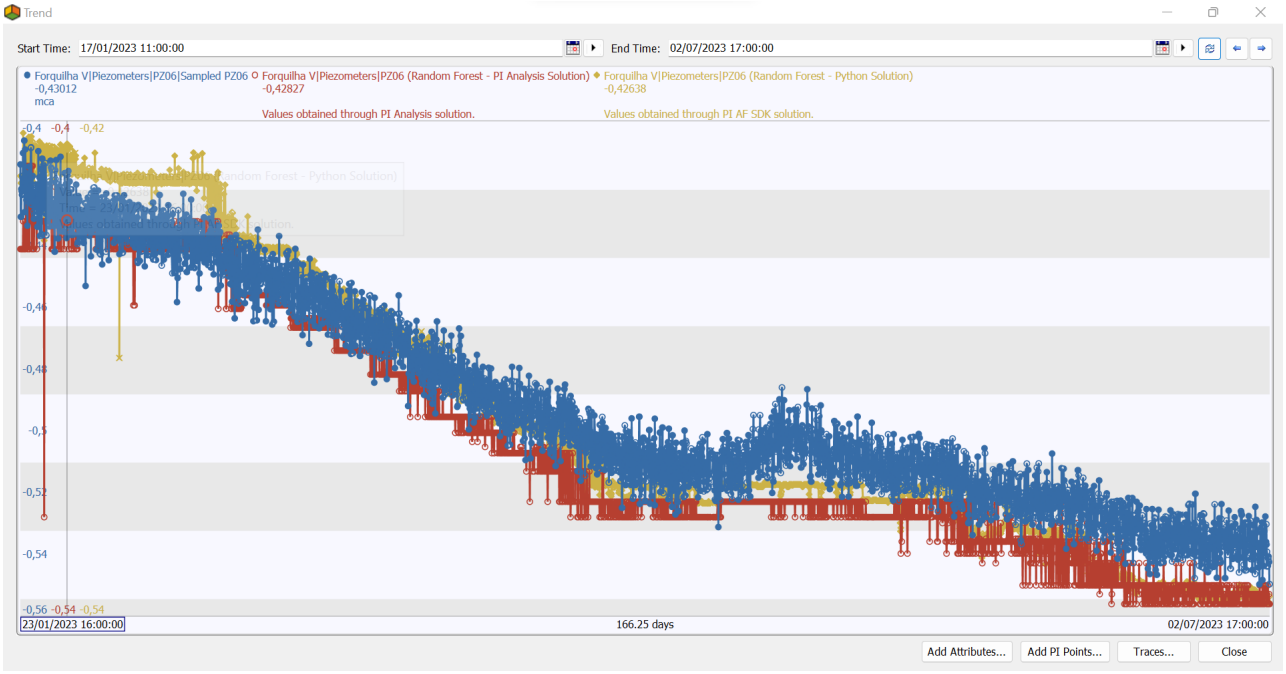


Fonte: Os Autores

6.3 Comparação entre os métodos de implementação

Avaliando o gráfico da Figura 6.6 notamos que a curva em vermelho, representando os valores obtidos pela solução implementada via PI Analysis se demonstrou mais ruidosa, com alguns outliers, demonstrando-se menos acertiva. Já a curva em amarelo, representando os valores obtidos pela solução implementada via Python demonstrou-se menos ruidosa e seguindo melhor a tendência da curva de valores esperados, representada pela tendência em azul.

Figura 6.6: Valores Esperados X Valores Obtidos - PZE06 Python Solution X Valores Obtidos - PZE06 PI Analysis Solution



Fonte: Os Autores

A tabela 6.1 nos mostra também que as métricas de avaliação RMSE, MAE e R^2 foram superiores para a implementação em Python:

Tabela 6.1: Métricas de performance: PI Analysis x Python

	MAE	**RMSE**	R^2
RF PI Analysis	0,004905	0,006102	0,985108
RF Python	0,002930	0,004231	0,985481

Fonte: Os Autores

Diante da análise exposta nota-se que os dois métodos de implementação apresentaram resultados satisfatórios, entretanto devido a complexidade do modelo, o método de implementação via Python demonstrou-se superior. A implementação da árvore via PI Analysis tornou-se complexa devido a sua profundidade e tamanho, sendo necessário simplificá-la, esse fato gerou perda de acurácia no modelo implementado. Conclui-se então que para modelos cuja implementação seja mais complexa recomenda-se a sua aplicação via Python devido a dificuldade de codificação no PI Analysis.

Capítulo 7

Conclusão

Com base na pesquisa e experimentação realizadas nesta dissertação, é possível concluir que a aplicação de técnicas de machine learning para a geração de sensores virtuais representa uma abordagem promissora e valiosa para a automação e o monitoramento de processos industriais, em especial para a instrumentação básica de barragens de rejeito, na qual os sensores físicos estão mais propensos a falhas, gerando anomalias e gap de dados. Ao testar vários métodos, incluindo MLP, Random Forest, Regressão Linear, Gradient Boosting e Decision Tree. Identificou-se para os dados analisados de piezômetros elétricos o Random Forest como o modelo mais eficaz.

Um aspecto notável desta pesquisa foi a exploração da correlação entre os múltiplos piezômetros estudados, que por sua vez revelou-se de grande importância para o estudo em questão. Constatou-se que a correlação fraca entre esses sensores pode impedir o aprendizado eficaz com os dados, essa descoberta é verdadeiramente importante e deve ser levada em consideração na aplicação de modelos de machine learning em situações semelhantes.

O estudo de janelas temporais, apesar de não demonstrarem de forma preliminar melhorias significativas nos modelos analisados, sugerem que a consideração do tempo passado deve ser melhor analisada para uma contribuição efetiva de memória no contexto de sensores virtuais em processos industriais. Nesse sentido deve-se haver um estudo aprofundado da dinâmica de cada instrumento avaliado, uma vez que um padrão específico pode não ser efetivamente capturado por uma janela de tempo limitada, não sendo portanto suficiente para fornecer informações relevantes para o modelo.

Após a seleção do modelo mais adequado, aprofundamos nossa análise, considerando diferentes configurações para o Random Forest, como o número de árvores e suas respectivas profundidades. Os resultados indicaram que não limitar a profundidade das árvores melhorou o desempenho do modelo. Isso fornece insights valiosos para a otimização futura do modelo.

Por fim, testes de normalidade e similaridade foram realizados para refinar ainda mais o modelo, garantindo a validade das previsões e a adequação à natureza dos dados,

permitindo desse modo a seleção de parâmetros significativos para o tempo de treinamento e recursos computacionais demandados. Ajustá-los permitiu encontrar um equilíbrio entre desempenho e eficiência.

Em última análise, esta dissertação contribui para o conhecimento no campo da geração de sensores virtuais e demonstra a importância da seleção criteriosa de modelos, ajuste de parâmetros, bem como da consideração da correlação entre as variáveis estudadas. Esperamos que este trabalho inspire pesquisadores e profissionais a explorar ainda mais o potencial dos métodos de machine learning em aplicações industriais e a desenvolver abordagens mais eficazes para a automação e o monitoramento de processos complexos.

Capítulo 8

Trabalhos Futuros

As barragens de rejeito estão equipadas com um número considerável de sensores de instrumentação, apesar da robustez desses devices, o ambiente hostil em que estão localizados pode afetar a qualidade de dados gerada por esses equipamentos, resultando em anomalias ou perdas. Este trabalho propós inicialmente a utilização de modelos de machine learning consolidados na literatura capazes de estimar sensores virtuais que podem auxiliar no processo de detecção de anomalias e preenchimendo de gaps em séries temporais fornecidas por piezômetros de corda vibrante. Os resultados inciais obtidos demonstraram que o modelo selecionado foi eficiente para piezômetros de corda vibrante, entretanto em uma próxima evolução deste trabalho pretende-se submeter o modelo mesclando outros tipos de sensores de instrumentação básica, com o objetivo de gerar variabilidade nos dados de entrada e observar como o modelo se comporta. Além disso pretende-se evoluir o modelo para aprendizado automático de máquina, de modo que o modelo possa se adaptar a novas amostras entrada de maneira autônoma.

Referências Bibliográficas

AVEVA, G. "**PI System Architecture, Planning and Implementation Course**", 2022. Disponível em: <http://cdn.osisoft.com/learningcontent/pdfs/PISystemArchitecturePlanningAndImplementationWorkbook.pdf>.

BADHIYE, S. S., CHATUR, P., WAKODE, B. "Data logger system: A Survey", ***International Journal of Computer Technology and Electronics Engineering (IJCTEE)***, pp. 24–26, 2011.

BISHOP, C. M., NASRABADI, N. M. ***Pattern recognition and machine learning***, v. 4. Springer, 2006.

BREIMAN, L. "Random forests", ***Machine learning***, v. 45, pp. 5–32, 2001.

BREIMAN, L. ***Classification and regression trees***. Routledge, 2017.

CARVALHO, F. B. D., TORRES, B. S., FONSECA, M. D. O., . "Sistemas PIMS-conceituação, usos e benefícios", ***Tecnologia em Metalurgia, Materiais e Mineração***, v. 1, n. 4, pp. 1–5, 2013.

DUNNICLIFF, J. ***Geotechnical Instrumentation for Monitoring Field Performance***. Canada: Wiley, 1988.

EHLERS, R. S. ***Análise de séries temporais***. Departamento de Estatística,UFPR, 2007.

FERNÁNDEZ-DELGADO, M., CERNADAS, E., BARRO, S., . "Do we need hundreds of classifiers to solve real world classification problems?" ***The journal of machine learning research***, v. 15, n. 1, pp. 3133–3181, 2014.

FONSECA, A. D. R. ***Auscultação por instrumentação de barragens de terra e enrocamento para geração de energia elétrica–Estudo de caso das barragens da UHE São Simão***. Tese de Doutorado, Dissertação de Mestrado, Programa de Pós-graduação da Universidade Federal . . . , 2003.

FORTUNA, L., GRAZIANI, S., RIZZO, A., . ***Soft sensors for monitoring and control of industrial processes***, v. 22. Springer, 2007.

FURQUIM, G., FILHO, G. P., JALALI, R., . "How to improve fault tolerance in disaster predictions: a case study about flash floods using IoT, ML and real data", ***Sensors***, v. 18, n. 3, pp. 907, 2018.

GAIOTO, N. ***Introdução ao projeto de barragens de terra e enrocamento***. São Carlos: Ed.USP São Carlos, Brasil, 2003.

GOODFELLOW, I., BENGIO, Y., COURVILLE, A. ***Deep learning***. MIT press, 2016.

GUIDE, B. "**Beginner's Guide**", 2002.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. H., . ***The elements of statistical learning: data mining, inference, and prediction***, v. 2. Springer, 2009.

JAMES, G., WITTEN, D., HASTIE, T., . ***An introduction to statistical learning***, v. 112. Springer, 2013.

MACHADO, W. G. F. ***Monitoramento de Barragens de Contenção de Rejeitos da Mineração***. Dissertação de mestrado - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Minas e Petróleo, 2007.

MARKLE FERNANDES VIEIRA, DILSON JUNIOR DE SOUSA LOPES, A. F. L. N. R. S. B. "INFLUÊNCIA DO LAGO DA UHE TUCURUÍ SOBRE A BARRAGEM DE CONCRETO: UM ESTUDO SOBRE OS MTJ'S", ***CIATEC-UPF***, v. 9, n. 10, pp. 14, 2017.

MÜLLER, S. H., FINDEISEN, R., ALLGÖWER, F. "Virtual sensors: A review", ***Annual Reviews in Control***, v. 49, pp. 141–157, 2020.

SCIENTIFIC, C., AS, I., WHOLE, A. "**CR300 Datalogger**", 2013.

SOARES, L. "Barragem de rejeitos". **CETEM/MCT**, 2010.

SOBREIRA, S. G. A. ***Aplicação de Sensores Virtuais Baseados em Aprendizado de Máquina para Estimativa de Vazão Mássica de Minério de Ferro em Correias Transportadoras***. Tese de Mestrado, Universidade Federal de Ouro Preto - Programa de Pós-Graduação em Instrumentação, Controle e Automação de Processos de Mineração, 2021.

SOBREIRA, S. G. A., GOMES, P. H., FILHO, G. P. R., . "A Data-Driven Soft Sensor for Mass Flow Estimation", ***IEEE Transactions on Instrumentation and Measurement***, v. 72, pp. 1–9, 2023. doi: 10.1109/TIM.2023.3273658.

UEYAMA, J., FAIÇAL, B. S., MANO, L. Y., . "Enhancing reliability in wireless sensor networks for adaptive river monitoring systems: Reflections on their long-term

deployment in Brazil", ***Computers, Environment and Urban Systems***, v. 65, pp. 41–52, 2017.

VARUN CHANDOLA, ARIDAM BANERJEE, V. K. ***Anomaly Detection: A Survey***. ACM Computing Surveys, 2009.

VIEIRA, A. C., GARCIA, G., PABÓN, R. E., . "Improving flood forecasting through feature selection by a genetic algorithm–experiments based on real data from an amazon rainforest river", ***Earth Science Informatics***, v. 14, pp. 37–50, 2021.

YAN, H., WANG, J., CHEN, J., . "Virtual sensor-based imputed graph attention network for anomaly detection of equipment with incomplete data", ***Journal of Manufacturing Systems***, v. 63, pp. 52–63, 2022.