UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Utilização de machine learning para prever atingimento de duas metas do marco legal do Saneamento Básico no Brasil

Alex Maluf Bastos

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Utilização de machine learning para prever atingimento de duas metas do marco legal do Saneamento Básico no Brasil

Alex Maluf Bastos

USP - São Carlos

2025

Alex Maluf Bastos

Utilização de machine learning para prever atingimento de duas metas do marco legal do Saneamento Básico no Brasil

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial.

Orientadora: Profa. Dra. Heloisa de A. Camargo.

USP - São Carlos

2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Maluf Bastos, Alex
M327u Utilização de machine learning para prever atingimento de duas metas do marco legal do Saneamento Básico no Brasil / Alex Maluf Bastos; orientadora Heloísa de Arruda Camargo. -- São Carlos, 2025.
52 p.

Trabalho de conclusão de curso (Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. Inteligência Artificial. 2. Aprendizado de Máquina. 3. Saneamento Básico. 4. Regressão Linear. 5. Random Forest. I. de Arruda Camargo, Heloísa, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

DEDICATÓRIA

A minha esposa e a meus filhos pela motivação e pela compreensão.

AGRADECIMENTOS

A minha orientadora Dra. Heloisa de A. Camargo, por direcionamento, suporte, correções e sugestões valiosas durante o processo de elaboração deste TCC.

Ao amigo Dr. Márcio de Lima e Silva pela avaliação, críticas e sugestões.

Agradeço também aos meus colegas de turma pelo compartilhamento de experiências, incentivo e pelas trocas de ideias ao longo do curso.

Aos meus pais, meu profundo agradecimento, por terem me dado o suporte familiar, o exemplo ético e a estabilidade emocional e financeira, vocês tornaram este percurso possível. Esta conquista também é de vocês, que acreditaram em mim e me apoiaram em cada passo do caminho.

A minha esposa e aos meus filhos, a minha mais profunda gratidão. A jornada do curso foi permeada pelo apoio de vocês. Em cada desafio, a resiliência de nossa família foi a força motriz para continuar.

EPÍGRAFE

" Onde há trabalho, não há milagre... Não falem de dons, de talentos inatos!...O gênio não é nada mais do que a diligência suprema."

Friedrich Nietzsche (2017)

RESUMO

BASTOS, Alex Maluf **Utilização de machine learning para prever atingimento de duas metas do marco legal do Saneamento Básico no Brasil**. 2025. 52 f. Monografia (MBA em Ciências de Dados) – Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

Este trabalho tem o objetivo de analisar a utilização de inteligência artificial para prever atingimento de duas das quatro metas do marco legal de 2020 do saneamento básico no Brasil. As metas escolhidas foram a meta de atendimento de 99% da população em abastecimento de água e 90% da população em coleta e tratamento de esgoto até 2033. Realizou-se a busca de dados sobre o andamento das obras e sobre a porcentagem de atendimento tanto em órgãos do governo quanto em entidades de monitoração da situação do saneamento básico no Brasil. Utilizou-se técnicas de ETL (extração, transformação e carga) de dados para obtenção de dados e o manuseio dos mesmos. Realizou-se posteriormente análise exploratória dos dados obtidos no intuito de entender relações entre as variáveis obtidas. Mediante a característica temporal dos dados, para realizar a previsão do atingimento das duas metas, foram escolhidas as metodologias de aprendizado de máquina chamadas regressão linear e *Random forest*. Ambas foram testadas, avaliadas e tiveram suas acurácias e eficiências comparadas. O resultado das fases de treino e teste de aprendizado de máquina não foram satisfatórios. Concluiu-se que dados de entrada podem ter sido insuficientes no quesito volume, ou as variáveis independentes não foram suficientes para que os algoritmos de aprendizado de máquina tivessem acurácia e precisão desejados.

Palavras-chave: Inteligência Artificial. Aprendizado de máquina. Saneamento Básico.

Regressão Linear. Random Forest.

ABSTRACT

BASTOS, Alex Maluf **Using Machine Learning to Predict the Achievement of Two Targets of Brazil's Sanitation Legal Framework**. 2025. 52 f. Monografia (MBA em Ciências de Dados) – Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

This work aims to analyze the use of artificial intelligence to predict the achievement of two of the four goals of Brazil's 2020 basic sanitation legal framework. The chosen goals were the target of providing water supply to 99% of the population and sewage collection and treatment to 90% of the population by 2033. Data on the progress of works and the percentage of service coverage was collected from both government agencies and entities monitoring the basic sanitation situation in Brazil. ETL (Extraction, Transformation, and Loading) techniques were used for data acquisition and handling. Subsequently, an exploratory data analysis was performed on the obtained data to understand the relationships between the variables. Given the temporal nature of the data, the machine learning methodologies chosen to predict the achievement of the two goals were linear regression and random forest. Both were tested, evaluated, and their accuracy and efficiency were compared. The results of the machine learning training and testing phases were not satisfactory. It was concluded that the input data might have been insufficient in volume, or the independent variables were not adequate for the machine learning algorithms to achieve the desired accuracy and precision."

Keywords: Artificial Intelligence. Machine Learning. Basic Sanitation. Linear Regression. Random Forest.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Elemento opcional. É composto de uma relação alfabética das abreviaturas e siglas utilizadas no texto seguido do seu significado.

ABNT	–	Associação Brasileira de Normas Técnicas
AI	–	Artificial Intelligence
AM	–	Aprendizado de Máquina
CSV	–	*Comma Separated Values*
IA	–	Inteligência Artificial
IBGE	–	Instituto Brasileiro de Geografia e Estatística
ML	–	Machine Learning
PIB	–	Produto Interno Bruto
PMI	–	Project Management Institute
SNIS	–	Sistema Nacional de Informações sobre Saneamento
SGBD	–	Sistema de Gerenciamento de Banco de Dados
SQL	–	Structured Query Language

SUMÁRIO

Sumário

1 Introdução

1.1 Contextualização do tema

A precariedade do saneamento básico no Brasil é alarmante. Segundo dados do Sistema Nacional de Informações sobre Saneamento, em 2024, apenas 56% contavam com coleta de esgoto e 52% do esgoto gerado recebia tratamento (Instituto Trata Brasil (ITB), 2024).

Estudos mostram que esta grave situação traz, tanto problemas de saúde para a população atingida, quanto dificuldades para geração de PIB das áreas afetadas (Teixeira, 2014).

1.2 Justificativa e motivação

Um país não pode dar condições mínimas de vida apenas para uma parte da população, uma nação deve fornecê-las a todos seus habitantes. Dados do IBGE de 2024, mostram que o PIB brasileiro foi de R$ 11,7 trilhões, com PIB per capita de R$47.802,02(IBGE, 2024). Um país dessa magnitude econômica pode e deve prover saneamento básico para todos seus habitantes.

Para sanar este flagelo, em 15 de julho de 2020, entrou em vigor o Novo Marco Legal do Saneamento Básico (Lei nº 14.026/2020). Esta lei estabeleceu meta de atendimento de 99% da população em abastecimento de água e 90% da população em coleta e tratamento de esgoto até 2033(BRASIL, 2020). Este marco legal também estabelece mecanismos para fiscalização e acompanhamento das obras através de regulamentação e definição de condições contratuais, para que os contratados para execuções das obras se comprometam com o atingimento da meta.

Devido aos impactos negativos em saúde pública e na economia das áreas afetadas, torna-se importante que nos asseguremos que a meta será atingida, para que saibamos quando estarão resolvidos esses problemas advindos da falta de saneamento básico. Vários fatores influenciam o andamento das obras e o atingimento da meta portanto, existe a possibilidade de não a atingir. As técnicas de machine learning podem auxiliar na previsão da data de conclusão das obras.

1.3 Problema de pesquisa e objetivo

1.3.1 Objetivo Geral

Criar modelo de machine learning utilizando algoritmos que façam previsão da porcentagem de população sem água potável e previsão da população sem coleta e tratamento de esgoto no ano de 2033 e comparar com a meta estabelecida de atendimento, que é de 99% (noventa e nove por cento) da população com água potável e de 90% (noventa por cento) da população com coleta e tratamento de esgotos até 31 de dezembro de 2033.

1.3.2 Objetivos específicos

Definido o objetivo principal deste trabalho, foram definidos os seguintes objetivos específicos

- Obter dados sobre andamento da implantação do marco legal.
- Realizar estudo de dados disponíveis que atendam a necessidade da previsão proposta.
- Criar e testar modelos de aprendizado de máquina para obter a previsão desejada.

2 Referencial teórico

A lei 14.026(BRASIL, 2020), que institui o marco legal do saneamento básico no Brasil, define saneamento básico como conjunto de serviços públicos, infraestruturas e instalações operacionais de:

a) abastecimento de água potável

b) esgotamento sanitário

c) limpeza urbana e manejo de resíduos sólidos

d) drenagem e manejo das águas pluviais urbanas

A mesma lei definiu as seguintes metas:

> "Art. 11-B Os contratos de prestação dos serviços públicos de saneamento básico deverão **definir metas de universalização que garantam o atendimento de 99%** (noventa e nove por cento) **da população com água potável e de 90%** (noventa por cento) **da população com coleta e tratamento de esgotos até 31 de dezembro de 2033**, assim como metas quantitativas de não intermitência do abastecimento, de redução de perdas e de melhoria dos processos de tratamento."

- Atrasos em obras públicas pode acontecer pelos seguintes motivos (Gallardo, 2018).
- Os aditivos contratuais para prorrogação de prazo, assim como os de alteração de valor.
- nos estudos de viabilidade técnica e econômica, que devem anteceder a elaboração do projeto básico de uma obra, observam-se falhas que impactam diretamente o prazo de execução.
- Projetos básicos imprecisos ou incompletos.
- A desconsideração de fatores climáticos.
- Durante a etapa da execução contratual também ocorrem fatores com potencial para atrasar a entrega da obra.
- Outra questão de mais difícil solução e que se sobressai na etapa de execução é a demora nos repasses de recursos financeiros pelos órgãos financiadores.
- Não se pode negar que, muitas vezes, as demandas políticas são responsáveis pela definição de cronogramas fictícios, elaborados na esperança de que a obra fique pronta antes desse ou daquele evento; resultando nas inaugurações de obras inacabadas, tão comuns no período eleitoral.

A implantação de obras de saneamento básico referenciada acima gera um grande volume de dados. Esta vastidão, somada à complexidade do conjunto de dados, nos coloca o desafio de como processá-los. Devido a estas características de grandeza e complexidade, a hercúlea tarefa de analisar os dados se tornou impossível de ser feita por humanos. Graças à evolução da tecnologia da informação, hoje temos ferramentas computacionais que oferecem a possibilidade da análise em questão em tempo satisfatório, com custos aceitáveis. A principal tecnologia que pode ser usada para esta análise é a inteligência artificial (IA).

2.1 Trabalhos relacionados

Segundo o Project Management Institute (PMI, 2021, p. X), um projeto é definido como:

> "*Um esforço temporário empreendido para criar um produto, serviço ou resultado único. A natureza temporária dos projetos indica um início e um fim para o trabalho do projeto ou uma fase do trabalho do projeto. Os projetos podem ser independentes ou fazer parte de um programa ou portfólio.*"

A gestão de projeto um projeto complexo como do saneamento básico envolve uma gama de recursos materiais, financeiros, logísticos e de pessoas. O volume de dados para gerir e controla-lo pode ser muito grande e complexo. A IA pode auxiliar a gestão de projetos de várias formas.

Segundo Barboza e Aguiar (2025), a inteligência artificial pode ser uma ferramenta para melhorar eficiência e reduzir custos em obras de engenharia civil.

Em Silva e Santos (2024) podemos ver uma abordagem do uso de IA como mecanismo melhoria de produtividade de projetos com uso de *chatbot*.

A literatura acima citada que usou IA em gestão de projetos, inspirou a iniciativa de usar machine learning, como ferramenta de gestão de prazos do projeto de saneamento básico.

2.2 Inteligência Artificial

Antes de falar em IA, porém, é importante salientar que a Tecnologia da Informação (TI) sempre teve como centro e propósito, os dados. TI tem o intuito de processar, armazenar e resguardar os dados úteis para aplicações ou softwares, de forma que estes dados estejam disponíveis e utilizáveis no momento em que são necessários.

No passado, devido às limitações tecnológicas, os computadores não conseguiam lidar com altos volumes de dados na velocidade necessária. Inovações habilitaram a tecnologia a

lidar com estes grandes volumes, o atualmente chamado Big Data. Este termo, criado pelo Analista do Gartner, Doug Laney em 2001 pode ser caracterizado pela sigla 3V, que são as seguintes características abaixo:

- Volume – geração de grandes volumes de dados.
- Velocidade – dados que são gerados em uma velocidade muito grande.
- Variedade – Tipos de dados vão além de caracteres e números, incluindo imagens e vídeos, por exemplo.

Podemos citar como exemplos de big data os dados gerados por redes sociais, dados gerados por sensores e ainda imagens de todos exames médicos de uma região específica.

O fato de haver possibilidade de lidar com o Big Data em tempo hábil, permitiu que o campo de IA florescesse, pois ele se alimenta e depende de grande volume de dados.

O termo inteligência artificial engloba atualmente um campo de conhecimento tão vasto que é difícil defini-lo em poucas palavras. Apesar dessa dificuldade, existe uma subárea deste campo chamada "aprendizado de máquina", do inglês "machine learning" (ML), o qual pode ser definido e entendido mais facilmente. Segundo Russell & Norvig, (2020):

> "*Machine learning* é um subcampo de IA que estuda a habilidade de melhorar desempenho baseado em experiência. Alguns sistemas de IA usam métodos de *machine learning* para atingir competência, mas alguns não os usam.

Para compreender o termo ML, primeiro precisamos definir o que significa aprender no contexto de IA. Em Mitchell (1997), define-se aprendizado vagamente, para incluir qualquer programa de computador que melhora seu desempenho em alguma tarefa através de experiência". Sua definição formal é:

> "Dizemos que um programa de computador aprende a partir da experiência E com relação a uma classe de tarefas T e a uma medida de desempenho P, se seu desempenho nas tarefas em T, conforme medido por P, melhora com a experiência E."

Ainda sobre ML, em Marsland (2015) é estipulado que

> "*Machine learning* é sobre fazer computadores modificarem e adaptarem suas ações para que essas ações se tornem mais acuradas, onde acurácia é medida por quão bem as ações escolhidas refletem as corretas. "

Os algoritmos de ML diferem de algoritmos de processamento de dados mais usados atualmente no que tange ao uso e ao volume dos dados. Um algoritmo de entrada de cadastro, por exemplo, aceita poucos dados de entrada de um novo cadastro e os insere em um banco de dados, que guarda permanentemente estas informações. Algoritmos de ML trabalham com altos volumes de dados para predizer dados de saída.

Segundo Friedman, Hastie e Tibshirani (2009), no cenário para aprender a partir de dados, temos uma medida de saída quantitativa(número) ou qualitativa(categoria) que queremos predizer, baseando em conjunto de features (características). Estas características são os dados de entrada. Os dados são divididos em conjunto de treinamento e conjunto de teste. O algoritmo aprende um modelo usando os dados de treinamento e utiliza o conjunto de teste para testar este modelo. Usando estes dados, criamos um modelo de previsão por meio de um algoritmo de aprendizado, ou learner(aprendiz), que nos permitirá prever o resultado para novos objetos não vistos. Um bom modelo é aquele que prevê com precisão esse resultado.

O aprendizado realizado por técnicas de ML pode ser dividido nas categorias supervisionado e não supervisionado.

No aprendizado supervisionado, são construídos modelos para prever ou estimar uma saída, baseando-se em dados de entradas que contêm *label* (etiqueta de classificação de dados de entrada). Por exemplo, são fornecidas várias imagens como entrada, algumas de cachorro, etiquetadas como Cachorro, e imagens de outros animais, etiquetadas como Outros. Dessa forma, o modelo pode aprender através dos *labels*, quais são fotos de cachorro e quais não são. Uma vez treinado o modelo, são fornecidas outras imagens de animais sem o label, para que o modelo identifique, ou classifique, quais são de cachorro e quais não são.

No aprendizado não supervisionado, são fornecidos dados de entrada, os quais não possuem labels. Neste caso, apenas dados de entrada são os insumos do algoritmo, os quais são usados para treinar o modelo para classificar em grupos.

O aprendizado supervisionado ainda pode ser classificado como um problema de regressão quando a saída é uma medida quantitativa a qual é representada por um número e se adequa ao objetivo deste documento (Friedman, Hastie e Tibshirani, 2009).

Dentro da categoria de aprendizado supervisionado de regressão, existem vários algoritmos à disposição dos profissionais da área que podem ser usados para estimar a data da conclusão das obras do saneamento básico. Para o problema alvo deste estudo, foram usados os métodos regressão linear e *Random forest*, ambos explicados abaixo.

2.2.1 Regressão linear

Em Morettin e Bussab (2023) temos a seguinte definição de regressão linear:

> "*A regressão linear é um modelo estatístico que assume que a relação entre a variável dependente e um conjunto de variáveis explicativas é linear. O modelo é estimado para prever ou explicar a média da variável dependente em função das variáveis independentes"*

Em James et al. (2013) regressão linear é definida como

> *"A regressão linear simples faz jus ao seu nome: é uma abordagem muito direta para prever uma resposta quantitativa Y com base em uma única variável preditora X. Ela pressupõe que existe uma relação aproximadamente linear entre X e Y."*

Na regressão linear simples pressupõe-se que os valores vêm de uma função matemática e, com os dados conhecidos, tenta inferir essa função. Uma vez descoberta, ela pode se usada para prever valores de y ao se supor um valor de x ainda desconhecido.

A regressão linear utiliza conjuntos de 2 variáveis, onde variável x (variável independente) determina o valor de variável y (variável dependente). Com estas informações, o algoritmo de machine learning consegue estimar a função que determina valores y a partir de valores de x. Desa forma, é possível utilizar os dados existentes para prever o valor de y desconhecido, a partir de um x. A regressão linear múltipla usa um conjunto de variáveis independentes para determinar a variável dependente, também chamada atributo alvo, ou label.

2.2.2 Random Forest

O método de aprendizado de máquina supervisionado chamado *decision tree,* ou árvore de decisão, pode ser usado tanto para problemas de regressão quanto de classificação. Neste modelo, o algoritmo utiliza um atributo inicial e divide os dados baseado em valores desse atributo num primeiro nível. Posteriormente, utiliza um próximo atributo para dividir os dados novamente baseado nos valores desse outro atributo e criar um outro nível da árvore de decisão, e assim sucessivamente. O objetivo deste mecanismo é encontrar os conjuntos de valores de cada atributo que levem a cada valor do atributo alvo. Uma vez descoberta a cadeia de valores de atributos que levem ao atributo alvo desejado, o algoritmo pode usar esta informação para prever valores de atributo alvo ainda desconhecido para valores de variáveis inseridas no modelo de aprendizado.

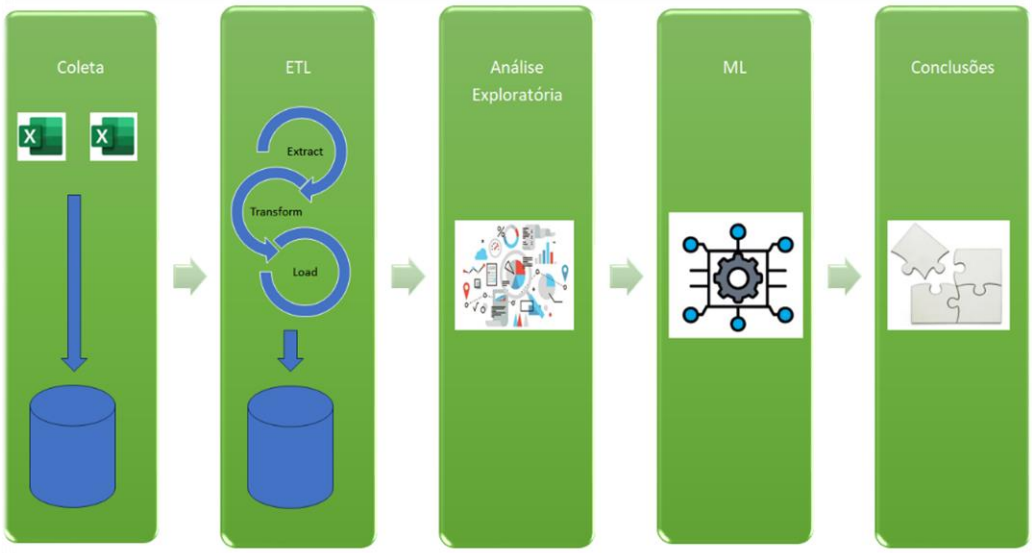
Random forest é do tipo comitê, composto por várias árvores de decisão com objetivo de criar modelo que consiga prever valores atributos para dados ainda desconhecidos.

3 Desenvolvimento

No presente documento, foi desenvolvida a proposta de uso de machine learning na averiguação de atingimento de meta de entrega dos itens abastecimento de água potável e esgotamento sanitário de saneamento básico. Esta proposta poderá ser utilizada para monitorar outros tipos de projetos de mesmo cunho.

O fluxo de trabalho adotado neste trabalho está descrito sucintamente abaixo.

Figura 1 – Fluxo de trabalho



- **Coleta de dados** – processos de coleta de dados em instituição, que divulga *status* do saneamento básico no Brasil.
- ***ETL (Extract, Trasnform and Load)*** – Após a coleta, seguiu-se processo de extração (**E**xtraction), transformação (**T**rasnform) e carga (**L**oad) de dados em um SGBD (sistema de banco de dados relacional) *Postgresql*.
- **Análise exploratória de dados** – Uma vez consolidados os dados em banco de dados relacional, foi realizada análise destes dados utilizando técnicas de exploração de ados para obter insights e conhecimento maior dos dados e o relacionamento entre estes mesmos dados. Nesta etapa optou-se por uso da linguagem de programação Python que tem ferramentas dedicadas a análises estatísticas de dados qualitativos e quantitativos.

- **Machine learning** – aplicaram-se mecanismos de *Machine learning* , ou aprendizado de máquina, para podermos prever se a meta do saneamento básico será cumprida.

As etapas do processo demonstrado acima serão detalhadas nos itens abaixo.

3.1 Coleta de dados

As informações do andamento das obras em questão estão disponíveis na internet, no entanto, obtê-los não foi um processo simples devido a algumas dificuldades. A primeira dificuldade se deve à distribuição de responsabilidade da implantação das obras. O governo definiu a meta no plano diretor, porém, as entregas das obras dependem das esferas de governos federal, estadual e municipal, o que implica em responsabilidade distribuída destas divisões na publicação de informações. Um segundo fator relevante sobre o acesso aos dados é que, apesar de públicos, muitos destes dados só são disponibilizados através de requisições. Apesar de ter sido feita a requisição, os dados não foram disponibilizados.

Diante do exposto, trabalhou-se com informações divulgadas sobre a evolução da cobertura de saneamento básico no país disponibilizada pela entidade Painel do Saneamento Básico.

Dados adicionais sobre população e PIB de unidade federal foram obtidos do IBGE, para enriquecer dados que poderiam ser fatores influentes no andamento das obras.

Os dados de andamento da obra obtidos se apresentaram em diversas planilhas anuais dos anos 2011 a 2022, no formato de planilha do Microsoft Excel, ainda não adequados para uma importação para banco de dados relacional. Segue exemplo do formato geral na figura 2.

Figura 2 – Formato dos dados originais.

	A	B	C	D	E	F	G
1	Painel Saneamento Brasil - www.painelsaneamento.org.br						
2	2021						
3	Localidade	Parcela da população total que mora em domicílios sem acesso à água tratada (% da população) (SNIS)	Parcela da população total que mora em domicílios sem acesso ao serviço de coleta de esgoto (% da população) (SNIS)	Volume de esgoto não tratado (água consumida - esgoto tratado) (mil m3) (SNIS)	Razão entre volume de esgoto tratado e volume de água consumida (%) (SNIS)	Internações por doenças associadas à falta de saneamento (Número de internações) (DATASUS)	Óbitos por doenças gastrointestinais infecciosas na população total (Número de óbitos) (DATASUS)
4	Brasil (0)	15,8%	44,2%	5.221.572,64	51,2%	143.578	1.660
5	Norte (Região) (1)	40,0%	86,0%	466.603,83	20,6%	28.697	181
6	Nordeste (Região) (2)	25,3%	69,8%	1.261.309,04	35,5%	64.355	634
23	Alagoas (UF) (27)	25,9%	82,1%	118.849,96	20,5%	1.903	22
24	Sergipe (UF) (28)	10,9%	69,5%	55.055,91	34,9%	852	18
36	Distrito Federal (UF) (53)	1,0%	8,2%	21.370,42	86,7%	2.221	8
37	Manaus (Região Metropolitana) (131)	6,4%	78,0%	90.462,07	17,6%	2.542	19
38	Belém (Região Metropolitana) (151)	35,4%	81,2%	61.405,80	6,5%	2.757	14
57	Goiânia (Região Metropolitana) (521)	7,4%	25,8%	59.218,69	57,6%	1.085	22
58	Ariquemes (Município) (110002)	15,3%	97,8%	3.328,76	2,3%	226	0

A Tabela 1 mostra uma descrição dos dados e seus formatos, os quais foram tratados na próxima etapa de ETL do projeto:

Tabela 1 – Exemplo de dados originais

Nome da coluna	**Formato do dado**	**Exemplo de dado**
Ano	Número	2015
Localidade	Distintos formatos para o país, para as unidades federais, para regiões metropolitanas e para os municípios.	Brasil (0) Norte (Região) (1) Manaus (Região Metropolitana) (131) Novo Airão (Município) (130320)
"Parcela da população total que mora em domicílios sem acesso à água tratada (% da população) (SNIS)"	Número real com ponto decimal representado por vírgula e o símbolo de porcentagem representado.	19,0%
"Parcela da população total que mora em domicílios sem acesso ao serviço de coleta de esgoto (% da população) (SNIS)"	Número real com ponto decimal representado por vírgula e o símbolo de porcentagem.	54,6%
"Volume de esgoto não tratado (água consumida - esgoto tratado) (mil m3) (SNIS)"	Número real com ponto decimal representado por vírgula e separador de milhares representado com ponto.	5.478.195,33
"Razão entre volume de esgoto tratado e volume de água consumida (%) (SNIS)"	Número real com ponto decimal representado por vírgula e o símbolo de porcentagem.	36,2%
"Internações por doenças associadas à falta de saneamento (Número de internações) (DATASUS)"	Número inteiro com separador de milhares representado com ponto.	603.623
"Óbitos por doenças gastrointestinais infecciosas na população total (Número de óbitos) (DATASUS)"	Número inteiro com separador de milhares representado com ponto.	2.903
"Rendimento do trabalho das pessoas que moram em residências com saneamento básico (R$ por mês) (IBGE)"	Número real com ponto decimal representado por vírgula e separador de milhares representado com ponto.	1.591,03
"Rendimento do trabalho das pessoas que moram em residências sem saneamento (R$ por mês) (IBGE)"	Número real com ponto decimal representado por vírgula e separador de milhares representado com ponto.	1.270,59

A análise dos dados originais deixou claro que seriam necessárias, no mínimo, as alterações listadas abaixo, que seriam feitas na próxima fase do projeto:

- **Formato de dados numéricos no geral**: devido ao fato de que usaríamos bibliotecas Python em outras fases, bibliotecas estas utilizam padrão americano de formato numérico, os pontos de milhar precisariam ser eliminados e os separadores decimais representados por vírgula substituídos por ponto.
- **Formato de dados em porcentagem**: em sistemas de informação como bancos de dados e linguagens de programação, os números não utilizam símbolos além do separador decimal. Sendo assim, texto que representavam números e continham o símbolo de porcentagem atrelado, deveriam ser tratados para remover o símbolo de porcentagem.
- **Nomes das colunas**: nomes longos e com espaço dificultariam o tratamento no banco de dados relacionais. Sendo assim, teriam que ser trocados por nomes de colunas mais sucintos e sem espaços em branco.
- **Dados dispersos em planilhas anuais**: a consolidação destes dados diretamente para o banco de dados relacionais poderia ser feita, no entanto, não haveria distinção por ano dentro do banco de dados.

- **Localidade**: Conforme descrito acima, as informações de dimensões diferentes (país, estado e municípios) estavam nesta mesma coluna.
- **Linhas excessivas**: as duas primeiras linhas da planilha (fonte da informação e ano) eram metadados das informações, mas não tinham utilidade para o trabalho.

Optou-se por tratar algumas questões acima diretamente em planilhas quando factível, mas a maioria das questões de tipos de dados e problemas de modelagem de dados foram tratadas dentro do banco de dados, pois este tipo de sistema permite tratar vários dados ao mesmo tempo com poucas linhas de comando da linguagem SQL (Structured Query Language) com eficiência e precisão.

3.2 ETL

As fases de extração, transformação e carga de dados podem ser realizadas em série, no entanto, algumas dificuldades que surgem em uma das fases faz com que seja necessário revisitar atividade realizada em fase anterior. Com o objetivo de clareza e objetividade, o processo é detalhado a seguir em sua forma sequencial.

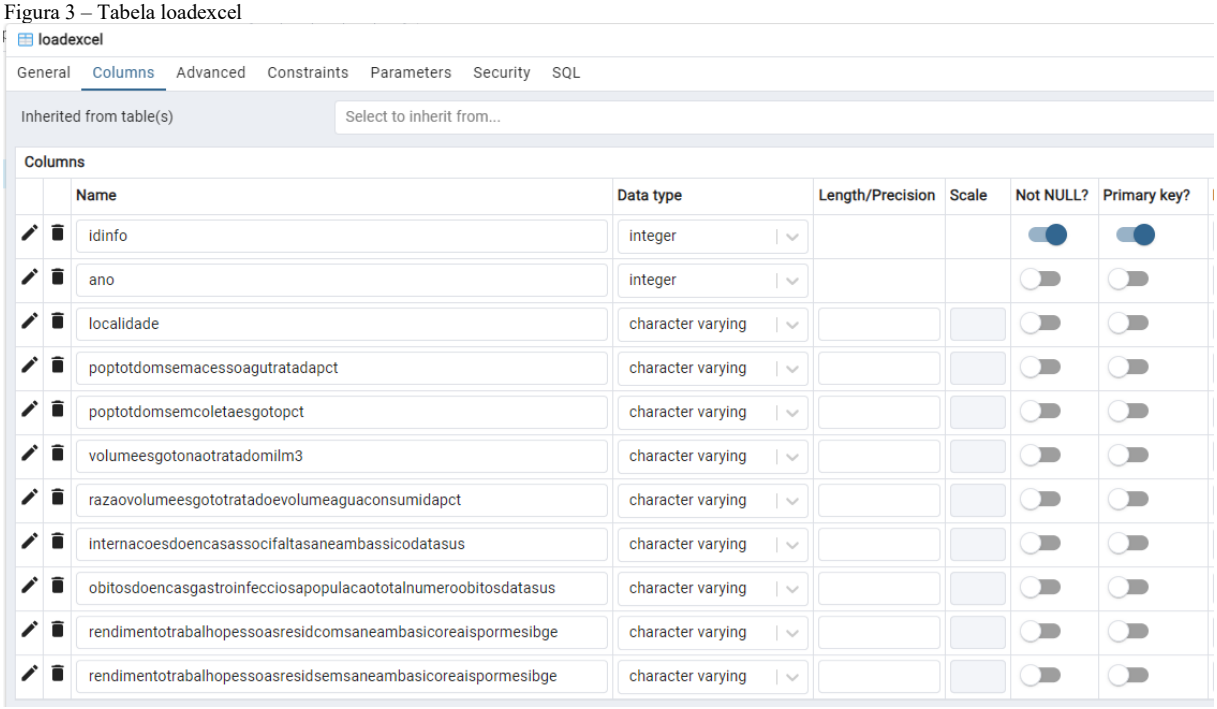
Transformação das planilhas Excel demandaram as seguintes ações:

- Inclusão da coluna ano em todas as planilhas para que fosse possível consolidá-las no banco de dados no momento da importação.
- Remoção das duas primeiras linhas pois estavam em um formato que não permitia que o restante dos dados fosse importado para banco de dados relacional.
- Cada planilha teve que ter dados úteis transpostos para formato CSV (Comma Separated Values), com valores separados por ponto e vírgula, pois é o formato que exigido pelo PostgreSQL para importação de dados. Os dados úteis não incluíam os nomes de coluna.

No PostgreSQL foram criadas as seguintes estruturas iniciais:

- Tabela **loadexcel**: destino das cargas dos arquivos CSV gerados anteriormente. Esta tabela possuía uma coluna com dado do tipo Integer configurada para ser a Primary Key da tabela. O restante das colunas correspondia às colunas dos arquivos CSV. Os nomes de colunas foram gerados a partir dos nomes de colunas das planilhas originais, porém, conforme explicado anteriormente, em formato mais adequado.

Figura 3 – Tabela loadexcel



- Tabela **dadointerm**: Esta tabela de dados intermediários foi desenhada para conter os dados consolidados de todas planilhas que seriam necessários para as próximas fases. Optou-se por dados desnormalizados nesta única tabela no intuito de facilitar as tratativas que viriam a ser feitas por Python na fase posterior. Nesta tabela também foram tratadas as questões impostas pela coluna localidade, através dos seguintes atributos ou colunas:
    - **tipodivisao**, que foi desenhada para categorizar os dados de localidade em P(País), R(Região), UF(Unidade Federal) e RM(Região Metropolitana). Desta forma, não seria necessário ter várias tabelas para cada nível de divisão territorial.
    - **Regiao**: para as regiões N(Norte), S(Sul), L(Leste), O(Oeste) e CO(Centro-Oeste)
    - **Coduf**: código da unidade federal, a ser usado pelos dados de nível unidade federal, município e região metropolitana.
- Outras colunas compatíveis com as colunas da tabela loadexcel, conforme abaixo

Figura 4 – Tabela dadointerm – Parte 1

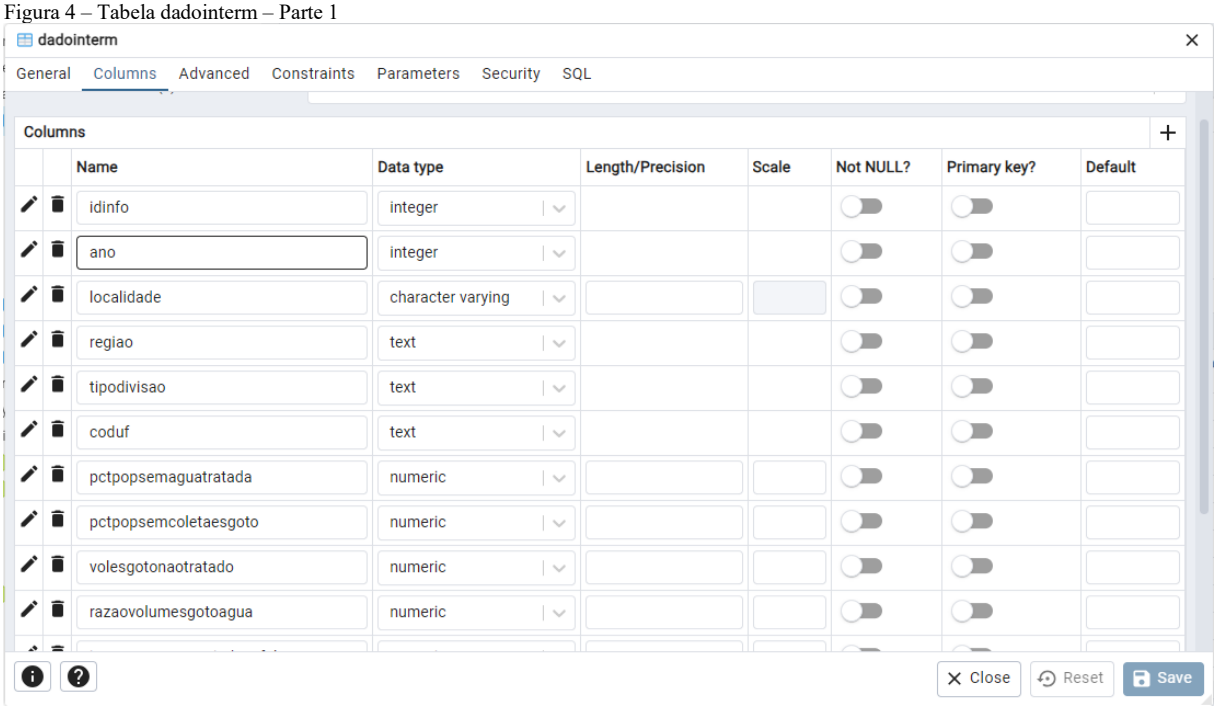
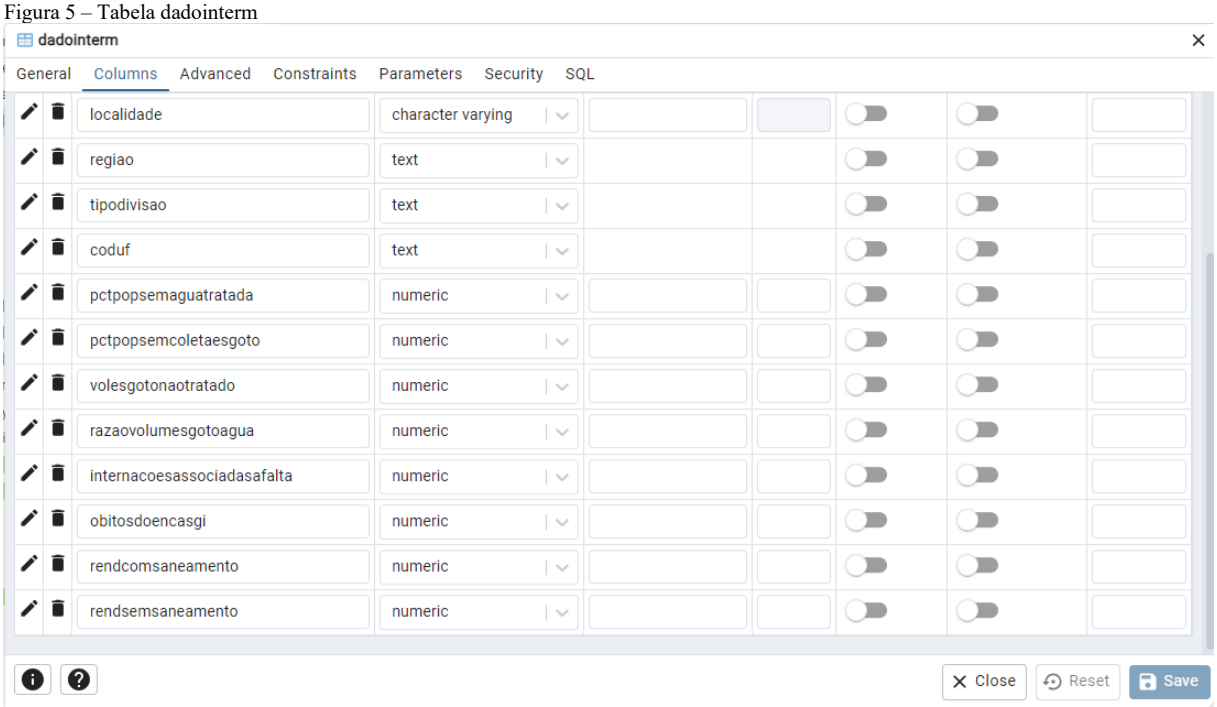


Figura 5 – Tabela dadointerm



- Tabela **ufregiao**: tabela utilizada para se relacionar com as outras tabelas com informações sobre código da região, nome da região, código de unidade federal, nome de unidade federal e abreviação de nome de região.

Conforme explanado acima, havia a necessidade de transformações de dados no formato de caracteres em formato numérico. Optou-se por realizá-las através de views de bancos de dados, que são representações parciais de dados de uma ou mais tabelas relacionadas. Usando-se views, não foi necessário alterar significativamente os dados para serem tratados por Python.

As views também contemplaram a informação de variação anual dos índices abaixo.

- "Parcela da população total que mora em domicílios sem acesso à água tratada (% da população) (SNIS - Sistema Nacional de Informações sobre Saneamento)"
- "Parcela da população total que mora em domicílios sem acesso ao serviço de coleta de esgoto (% da população) (SNIS - Sistema Nacional de Informações sobre Saneamento)"

Dessa forma, possibilitou-se usar estes dados nos modelos de machine learning utilizados na próxima fase deste trabalho.

A princípio, foram criadas 3 *views* por divisão territorial, sendo uma para região, uma para unidade federal e uma por município.

As *views* adicionaram as seguintes colunas:

- Deltapctpopsemaguatratada: Variação anual da porcentagem de população sem água tratada
- Deltapctpopsemcoletaesgoto: Variação anual da porcentagem de população sem coleta de esgoto.

No código para estes deltas, foram utilizadas window function lag. O primeiro ano para cada localidade foi configurado como zero.

Ao final da transformação e da análise exploratória dos dados feita posteriormente, a *view* com dados por unidade federal que foi usada como matriz de dados de entrada dos modelos de aprendizagem possuía as seguintes colunas.

Tabela 2 – Descrição de colunas da view

Nome da coluna	**Significado**
ano	ano
localidade	localidade
ufsimples	apenas nome da unidade federal
pctpopsemaguatratada	Parcela da população total que mora em domicílios sem acesso à água tratada (% da população) (SNIS)
deltapctpopsemaguatratada	Diferença entre o ano e o ano anterior de "Parcela da população total que mora em domicílios sem acesso à água tratada (% da população) (SNIS)"
populsemaguatratada	População em números absolutos de "Parcela da população total que mora em domicílios sem acesso à água tratada "
pctpopcomaguatratada	Parcela da população total que mora em domicílios com acesso à água tratada
populcomaguatratada	População em números absolutos de "Parcela da população total que mora em domicílios com acesso à água tratada "
pctpopsemcoletaesgoto	"Parcela da população total que mora em domicílios sem acesso ao serviço de coleta de esgoto (% da população) (SNIS)"
deltapctpopsemcoletaesgoto	Diferença entre o ano e o ano anterior de "Parcela da população total que mora em domicílios sem acesso ao serviço de coleta de esgoto (% da população) (SNIS)"
populsemcoletaesgoto	População em números absolutos de "Parcela da população total que mora em domicílios sem acesso ao serviço de coleta de esgoto (% da população) (SNIS)"
pctpopcomesgototratado	"Parcela da população total que mora em domicílios com acesso ao serviço de coleta de esgoto (% da população) (SNIS)"
populcomcoletaesgoto	População em números absolutos de "Parcela da população total que mora em domicílios com acesso ao serviço de coleta de esgoto (% da população) (SNIS)"
populacao	População da localidade
deltapopulacao	Diferença entre ano e ano anterior da população
variacaopctpopulacao	Variação percentual anual da populaçao
pib	PIB da localidade
deltapib	Variação anual do PIB da localidade
variacaopctpib	Variação percentual anual do PIB da localidade
razaovolumesgotoagua	"Razão entre volume de esgoto tratado e volume de água consumida (%) (SNIS)"
internacoesassociadasafalta	"Internações por doenças associadas à falta de saneamento (Número de internações) (DATASUS)"
obitosdoencasgi	"Óbitos por doenças gastrointestinais infecciosas na população total (Número de óbitos) (DATASUS)"
rendcomsaneamento	"Rendimento do trabalho das pessoas que moram em residências com saneamento básico (R$ por mês) (IBGE)"
rendsemsaneamento	"Rendimento do trabalho das pessoas que moram em residências sem saneamento (R$ por mês) (IBGE)"
coduf	Codigo da Unidade Federal (IBGE)
codregiao	Código da região(IBGE)
regiao	Nome da região
abrevregiao	Abreviatura da região
tipodivisao	Tipo de divisão.

3.3 Análise exploratória de dados

Nesta fase, foi possível realizar aprofundamento nos valores dos atributos dos dados utilizando medidas estatísticas centrais, de momento e de distribuição.

Figura 6- Resumo dados por região(N, S, L, O, CO)

	ano	pctpopsemaguatratada	deltapctpopsemaguatratada	pctpopsemcoletaesgoto	deltapctpopsemcoletaesgoto
count	65.000000	65.000000	65.000000	65.000000	65.000000
mean	2016.000000	20.090769	-0.326154	58.484615	-0.958462
std	3.770776	13.415112	1.082139	23.774896	1.133564
min	2010.000000	8.200000	-3.800000	18.300000	-4.800000
25%	2013.000000	9.300000	-0.800000	42.300000	-1.700000
50%	2016.000000	11.800000	-0.200000	57.900000	-1.000000
75%	2019.000000	27.400000	0.100000	77.800000	-0.200000
max	2022.000000	47.600000	2.800000	93.500000	2.700000

Figura 7 - Resumo dados por UF(unidade federal)

	ano	pctpopsemaguatratada	deltapctpopsemaguatratada	pctpopsemcoletaesgoto	deltapctpopsemcoletaesgoto
count	351.000000	351.000000	351.000000	351.000000	351.000000
mean	2016.000000	23.782621	-0.373504	66.240456	-1.038177
std	3.746999	16.479284	2.839881	24.716195	3.000604
min	2010.000000	0.600000	-14.500000	6.300000	-20.900000
25%	2013.000000	13.200000	-1.200000	51.400000	-2.100000
50%	2016.000000	18.800000	-0.200000	73.900000	-0.700000
75%	2019.000000	29.000000	0.500000	85.400000	0.000000
max	2022.000000	67.100000	19.300000	97.900000	19.900000

Figura 8 - Resumo dados por município

	ano	pctpopsemaguatratada	deltapctpopsemaguatratada	pctpopsemcoletaesgoto	deltapctpopsemcoletaesgoto
count	11180.000000	10829.000000	10747.000000	10661.000000	10413.000000
mean	2016.000000	19.574938	-0.293608	57.933487	-0.851282
std	3.741825	23.410686	6.357773	37.007304	9.344402
min	2010.000000	-7.700000	-100.000000	-7.000000	-100.000000
25%	2013.000000	1.800000	-0.900000	20.000000	-1.200000
50%	2016.000000	10.000000	0.000000	64.500000	0.000000
75%	2019.000000	29.300000	0.200000	97.700000	0.000000
max	2022.000000	100.000000	100.000000	100.000000	100.000000

As análises de medidas centrais e percentis levou às seguintes descobertas:

- Volume de dados sobre regiões era muito menor do que volume de dados sobre unidade federal.
- No nível de município, houve anos em que o percentual de população atendida por água tratada diminuiu em relação ao ano anterior. A hipótese que se cogitou foi que a população cresceu e o atendimento de saneamento básico não acompanhou esse crescimento naqueles anos de variação negativa.

- No nível de município, dados de valores negativos de percentual de população atendida por água tratada, o que é incoerente. Para níveis de região e unidade federal, não havia esta inconsistência.

Após as análises acima, decidiu-se usar dados do nível de unidade federal pelos seguintes motivos:

- O nível de região possuía volume pequeno de dados (65 linhas), enquanto no nível de unidade federal havia um volume maior (351 linhas).
- Dados inconsistentes do nível de município.

Decidido o uso de dados nível de unidade federal, foram feitas análises adicionais nos dados.

Iniciou-se com uma comparação entre os estados utilizando boxplot de todas unidades federais. A figura 9 mostra esta comparação para ”porcentagem da população sem água tratada” e a figra10 no mesmo gráfico, mostra esta comparação para ”porcentagem da população sem coleta de esgoto”. Esta visualização mostrou que existem discrepâncias entre o atendimento da população para os itens analisados (água tratada e coleta de esgoto), onde pudemos observar que a situação de falta de coleta de esgoto é pior do que a situação de acesso a água tratada. Nota-se também que existem diferenças na variação anuais dos percentuais entre as unidades federais.

Figura 9 - *Boxplots* comparando UF no quesito “sem acesso a água tratada”

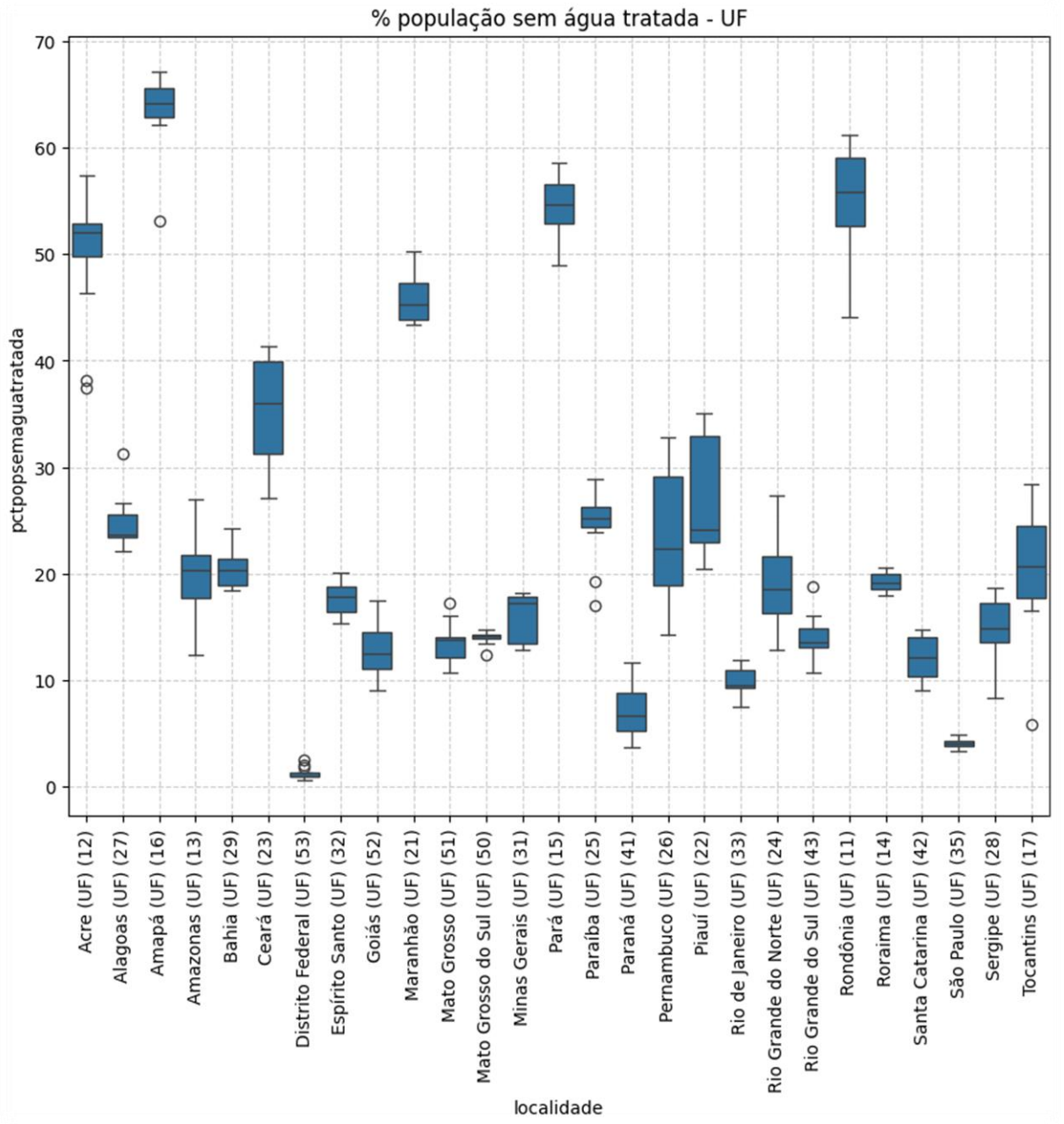
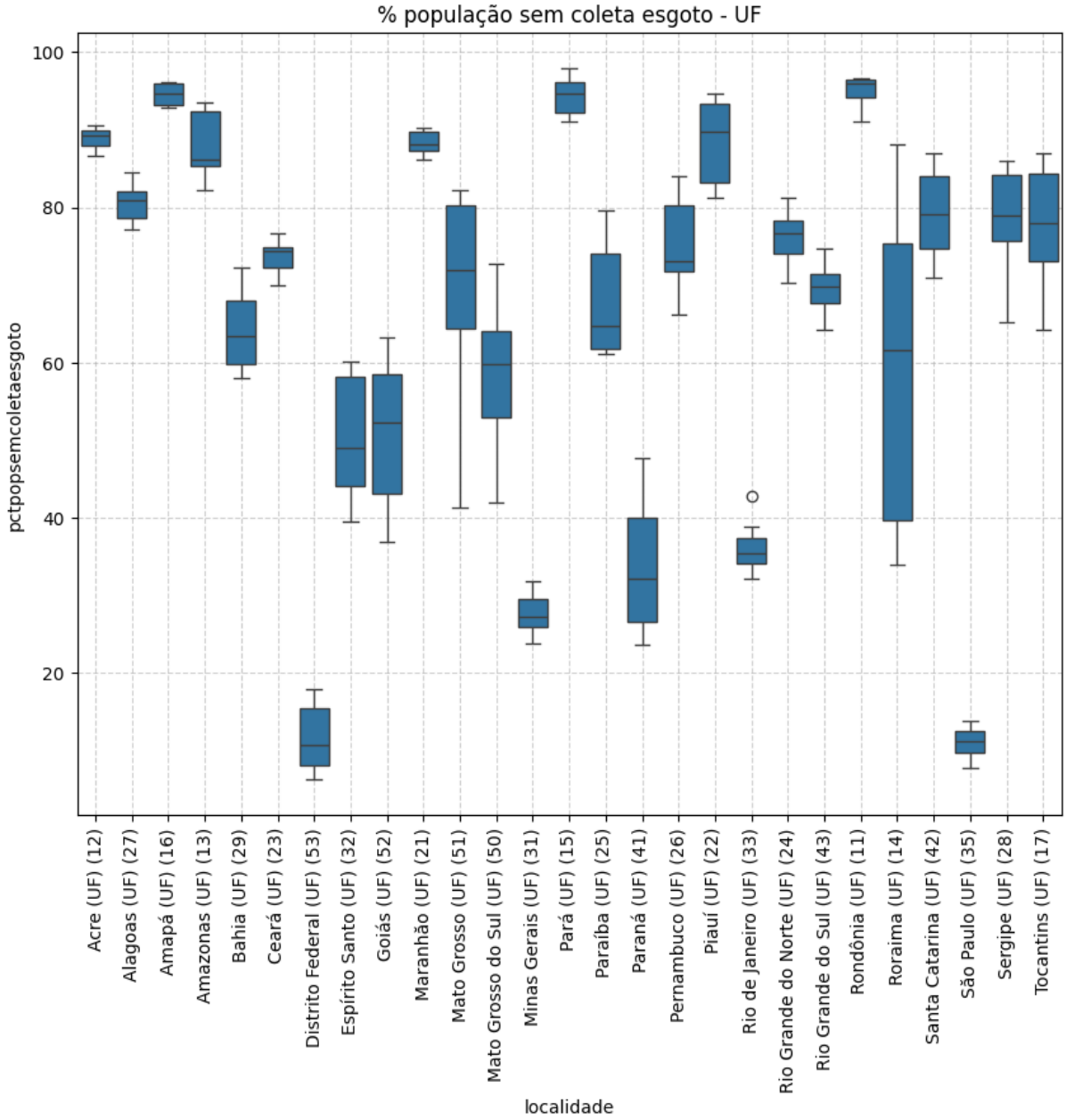
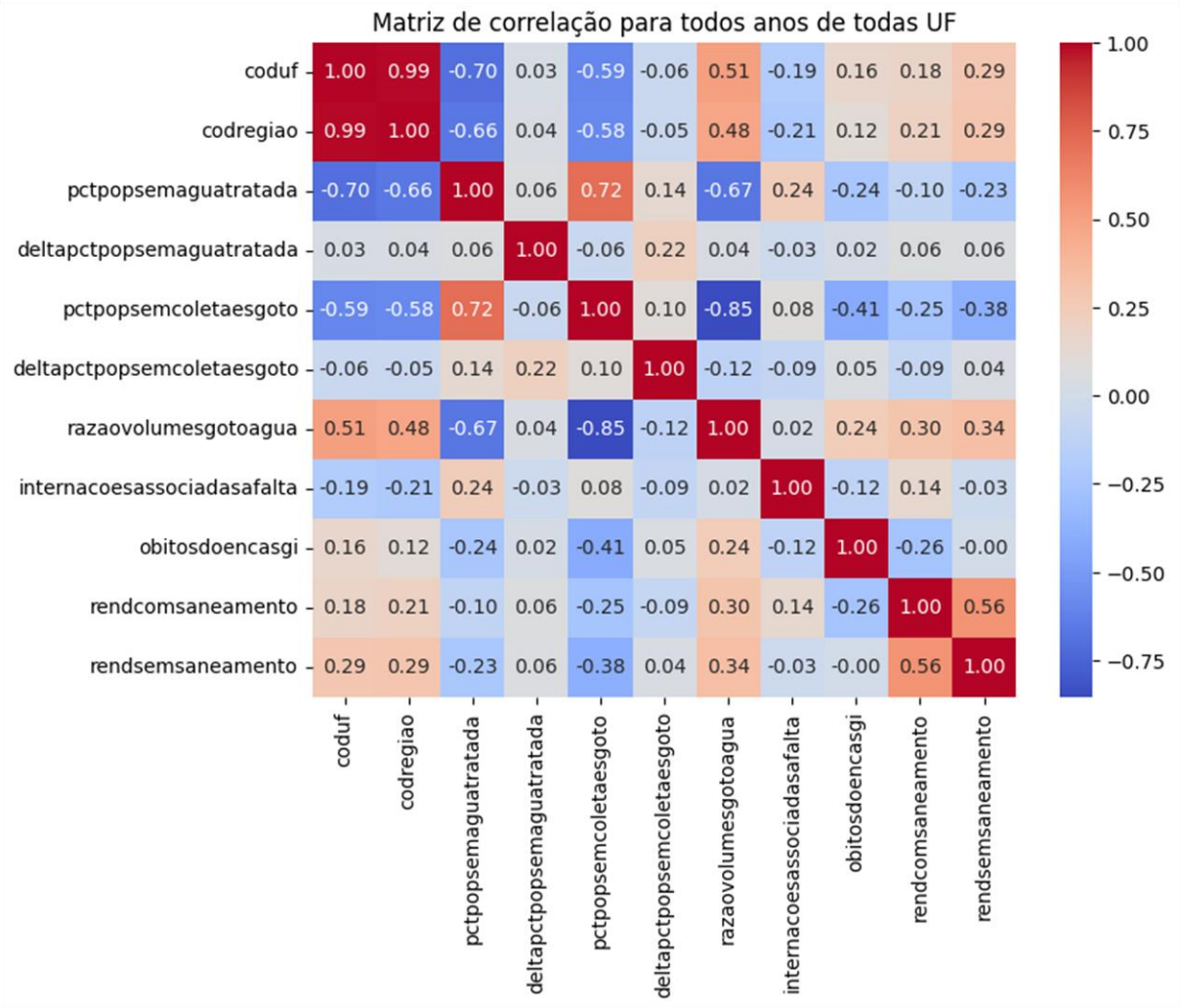


Figura 10 - *Boxplots* comparando UF no quesito “sem acesso a coleta de esgoto”



A análise seguinte focou na correlação dos atributos. A lista e significado das colunas da view estão descritos na Tabela2 deste documento. Como os atributos originais possuíam denominações extensas, ao se transformar esses dados no banco de dados, eles foram renomeados conforme foi explicado na seção de ETL, tabela 2.

Figura 11 – Primeira matriz de correlação



As variáveis da tabela mostram consequências da falta de saneamento básico, portanto, não foram utilizadas como insumo para análise de predição de finalização das obras.

Tabela 3 – Variáveis não usadas na fase de aprendizado de máquina.

Variável	Significado
razaovolumesgotoagua	"Razão entre volume de esgoto tratado e volume de água consumida (%) (SNIS)"
internacoesassociadasafalta	"Internações por doenças associadas à falta de saneamento (Número de internações) (DATASUS)"
obitosdoencasgi	"Óbitos por doenças gastrointestinais infecciosas na população total (Número de óbitos) (DATASUS)"
rendcomsaneamento	"Rendimento do trabalho das pessoas que moram em residências com saneamento básico (R$ por mês) (IBGE)"
rendsemsaneamento	"Rendimento do trabalho das pessoas que moram em residências sem saneamento (R$ por mês) (IBGE)"

A primeira matriz de correlação mostrou o seguinte:

- Correlação negativa de 0,70 entre porcentagem da população sem acesso a água tratada e unidade federal. Sendo assim, concluiu-se que existe relação inversa entre unidade federal e entrega de obras destes itens do saneamento básico. Não se pôde inferir a causalidade desta

correlação apenas com os dados usados. Isto fez com que houvesse busca por outras informações sobre as unidades federais que poderiam influenciar nos indicadores.

- Correlação positiva de 0,72 entre porcentagem de população sem acesso a água tratada e porcentagem da população sem acesso a coleta de esgoto. Esta relação não é causal, pois ambos indicadores têm a mesma causa, que é a ineficiência do governo em prover ambos. O mesmo acontece com as correlações com o índice de razão do volume de esgoto com água consumida.

A primeira correlação citada acima levou ao questionamento sobre quais seriam outras diferenças mais díspares entre os estados do Brasil, e decidiu-se usar os parâmetros de população e PIB como novas variáveis independentes que poderiam influenciar nos indicadores analisados. Sendo assim, foram realizadas as seguintes coletas adicionais de dados:

- Dados de PIB anual por unidade federal.
- Dados de população anual por unidade federal. Nesta coleta, foram necessárias interpolações de dados nos anos de 2014 e 2021, devido à ausência destes dados da fonte do IBGE.

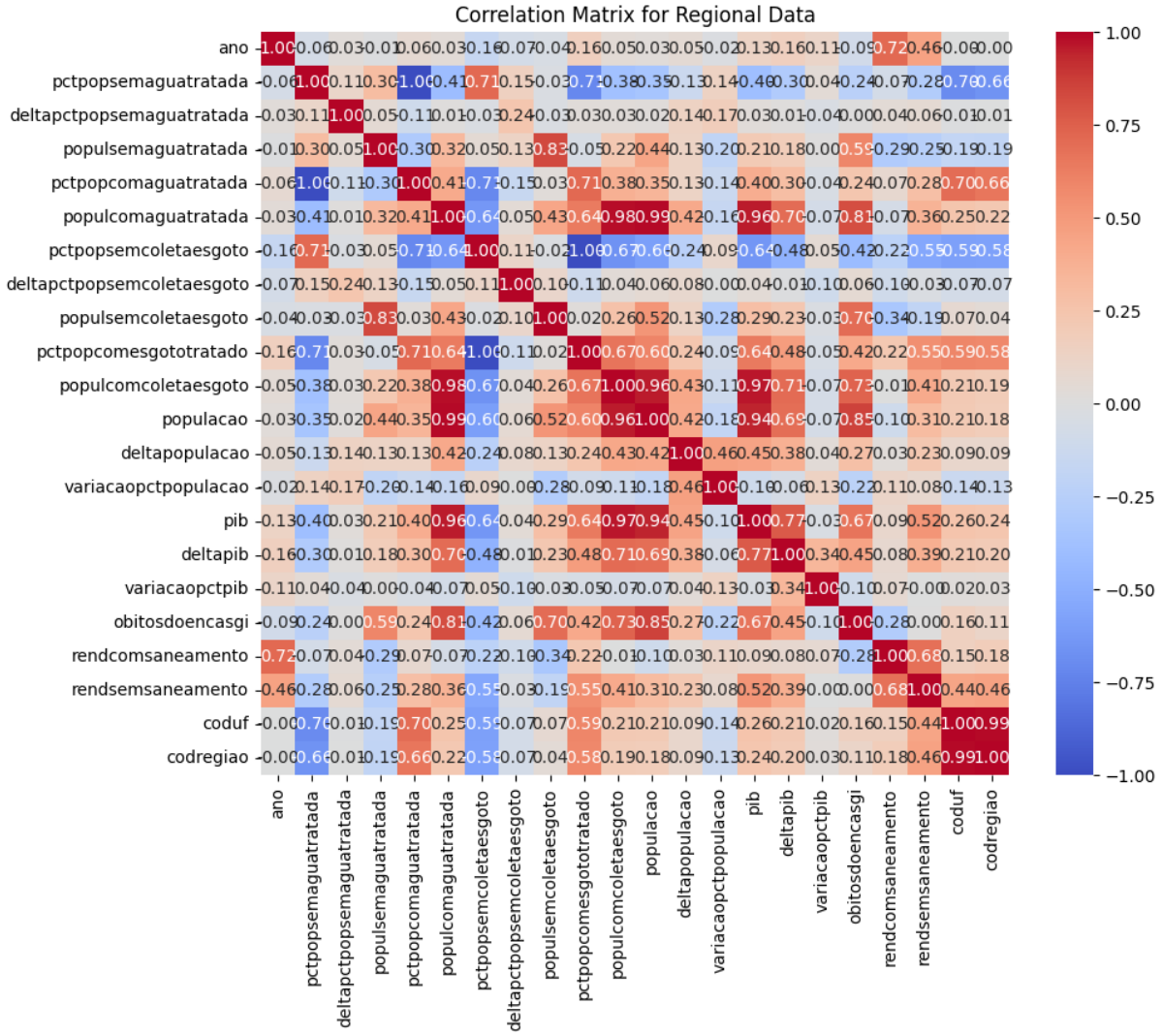
A adição de dados de PIB e população levou a mudanças nas views do banco de dados para que estes dados fossem apresentados em conjunto com os dados já analisados anteriormente. Além destes dados puros, foram adicionadas as seguintes colunas na view utilizando-se window function na query da view:

- Variação anual do PIB por unidade federal.
- Variação anual da população por unidade federal.
- População com água tratada.
- Porcentagem da população com água tratada.
- Variação anual de população sem água tratada.
- População com coleta de esgoto.
- Porcentagem da população com coleta de esgoto.
- Variação anual de população com água tratada.

Analisou-se então a questão de covariância entre todo os dados obtidos para analisar o que influenciaria os índices de população atendida por água encanada e por tratamento de esgoto.

Uma nova matriz de correlação foi gerada, conforme abaixo.

Figura 12 – Segunda Matriz de Correlação



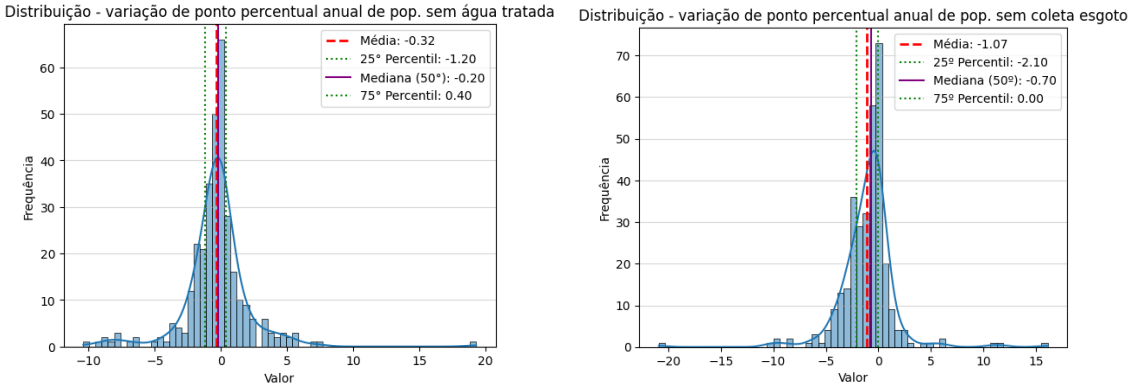
A inclusão de novos dados levou às seguintes observações:

- Correlação positiva de 0,96 entre o PIB com população com água tratada e também com população com tratamento de esgoto.
- Correlação positiva de 0,70 entre variação anual de PIB população com água tratada e também com população com tratamento de esgoto.
- Correlação positiva de 0,69 entre população e variação anual do PIB.

As correlações acima indicam que há possível influência do PIB e do tamanho da população na porcentagem de pessoas sem água tratada e sem coleta de esgoto.

Uma última análise foi realizada para que pudéssemos observar a distribuição do valor de variação anual da porcentagem de população não atendida pelos indicadores mostrados na figura 13.

Figura 13 – Distribuição de valores delta



A medida central média para ambos os indicadores analisados mostra que, como a variação é negativa, as obras de saneamento relativa a ambos estão avançando, o que diminui a porcentagem de população sem água tratada ou sem coleta de esgoto.

A variação de ponto percentual de população sem água tratada está concentrada entre 1,20 negativo e 0,4 positivo, resultando em IQR de -1,6, enquanto para ausência de coleta de esgoto, o IQR é de (0-(-2,10)) 2,1. Ambos gráficos exibem alguns outliers, o que era esperado pois o gráfico de *boxplots* mostrado anteriormente já denotava disparidades entre as unidades federais no tocante à porcentagem de população atendida por água tratada e coleta de esgoto. Estes histogramas evidenciaram também que não são regulares ao longo dos anos as variações do atributo de porcentagem de população não atendida por água tratada e coleta de esgoto.

3.4 Machine Learning

Após a análise exploratória dos dados, os atributos independentes escolhidos para serem usados na modelagem foram os seguintes:

- Ano
- Variação anual de pontos percentuais de população não atendida por água tratada ou coleta de esgoto.
- Variação anual de PIB da unidade federal.

Para o a variável dependente (label, ou target) foram usados dois atributos, o percentual da população sem água tratada e o percentual de população sem coleta de esgoto.

Os dados coletados apresentavam os atributos alvo entre os anos 2011 a 2022, e queríamos prever os valores destes atributos para o ano de 2033. Como atributos alvo faziam parte dos dados analisados, escolheu-se modelos de aprendizado supervisionado.

Para cada um dos indicadores (não acesso a água tratada e não acesso a coleta de esgoto), foram testados ambos modelos escolhidos. Visto que que ~~estes~~esses indicadores ~~são~~representam ausência de acesso a recursos~~, portanto~~ essenciais, espera-se que essas porcentagens diminuam ao longo do tempo. Para o PIB, por outro lado, espera-se que ele apresente crescimento contínuo e uma variação anual seja positiva.

Dada a variação temporal dos atributos nos dados coletados, optou-se inicialmente por testar a regressão linear entre os modelos de Aprendizado de Máquina (AM). No entanto, este modelo pressupõe que as variações dos atributos são regulares, premissa que não foi confirmada pela análise exploratória de dados. Diante disso, e visando comparar a eficiência com um algoritmo mais adequado a variações irregulares ao longo do tempo, foi selecionado também o Random Forest para a modelagem.

A linguagem de programação escolhida foi Python, devido à riqueza de bibliotecas disponíveis para estatística e para aprendizado de máquina.

Ambos os modelos escolhidos oferecem a possibilidade de se utilizar os mesmos atributos independentes para se prever o label desejado. O que os diferencia são os parâmetros de cada algoritmo.

Como o repositório de dados foi um SGBD PostgreSQL, foi necessário transformar os dados em formato adequado às bibliotecas de Python. Optou-se por usar a biblioteca Pandas, que foi utilizada para transformar os dados em *dataframes* do Pandas.

Para a modelagem de aprendizado de máquina, a escolha recaiu sobre a biblioteca *Scikit-learn*[1] pois ela possui vários módulos de modelos de IA, inclusive regressão linear e Random Forest, que foram aqueles escolhidos para este trabalho.

Os valores escolhidos para previsão foram:

- Ano: 2033
- Variação anual de pontos percentuais de população não atendida por água tratada ou coleta de esgoto: foi escolhida a média dos valores históricos, -10,4
- Variação anual de PIB da unidade federal: Foi escolhida a média dos valores históricos, 8,88

Foram seguidos os procedimentos de processo de aprendizado de máquina, constituído das seguintes fases:

- Treino: nesta etapa, dados conhecidos são usados para treinar um modelo. De todos dados disponíveis, foram usados 70% dos dados nesta etapa. Foram escolhidos os anos entre 2011 e 2018 para treino.
- Teste: nesta fase, o modelo já treinado usa diferentes dados daqueles usados em treinamento para averiguar a eficácia de aprendizagem do modelo. Neste trabalho foram usados 30% dos dados conhecidos nesta fase. Este conjunto de dados foram os anos entre 2019 e 2021 para treino.
- Execução: neste ponto, o modelo já treinado é testado é utilizado com dados não conhecidos para fazer a previsão desejada.
- Avaliação: eficiência dos algoritmos é avaliada através de indicadores dos cada modelo, comparando-se previsões dos dados reais de 2019 a 2021 com os dados reais destes [1]mesmos anos. Foram utilizados os seguintes indicadores de avaliação
    - Erro Absoluto Médio (MAE-Mean Absolute Error): quanto menor este valor, melhor a precisão da previsão
    - Erro Quadrático Médio (MSE-Mean Squared Error): quanto menor este valor, melhor a precisão da previsão
    - R2 score: o melhor valor é um.

As tabelas 5 e 6 mostram as medidas de avaliação da eficiência dos algoritmos, ao se comparar os resultados da fase de teste com os dados esperados.

[1] https://scikit-learn.org/stable/index.html

Tabela 5 – Avaliação de regressão linear

Modelo	Avaliação da regressão linear do indicador "**sem acesso a água tradada"**	Avaliação da regressão linear do indicador "**sem acesso a coleta de esgoto"**
Erro Absoluto Médio (MAE)	13,039	18,1937
Erro Quadrático Médio (MSE):	287,79	512,50
R2 score	-0,02	0,05

Tabela 6 – Avaliação de Random forest

Modelo	Avaliação da *random forest* do indicador "**sem acesso a água tradada"**	Avaliação da *random forest* do indicador "**sem acesso a coleta de esgoto"**
Erro Absoluto Médio (MAE)	16,1973	22,2886
Erro Quadrático Médio (MSE):	353,72	752,56
R2 score	-0.25	-0,19

O atributo alvo é uma porcentagem que pode ter valores de 0 a 100 %, e para ambos os modelos utilizados, MAE se mostrou muito alto, representando erros percentuais grandes, indicando uma margem de erro muito grande na previsão. O fato de a raiz quadrada do MSE ser maior que o MAE indica que alguns erros de previsão são consideravelmente maiores do que a média. O R2 score muito baixo indica que o modelo consegue explicar apenas porcentagem muito baixa da variância no total da porcentagem que queríamos prever. Esse valor indica que o modelo não tem utilidade preditiva.

Os mecanismos de avaliação dos dois modelos escolhidos apresentaram o mesmo comportamento, fato que levou à conclusão de que não adiantaria testar novos modelos com os mesmos dados, pois os dados coletados pareceram insuficientes para realização da previsão desejada.

4 Conclusão

A avaliação da eficiência dos modelos de aprendizagem de máquina mostrou que não foi possível fazer a previsão desejada do cumprimento de metas.

Obras públicas de engenharia civil de nível nacional, como as do saneamento básico, envolvem muitos fatores que determinam a eficiência e efetividade do andamento do projeto. A limitação de dados coletados pode ter influenciado nos fracos resultados das previsões dos algoritmos utilizados.

Acredito que uma análise que utilizasse mais dados existentes sobre as obras poderia levar o resultado mais assertivos e confiáveis dos modelos de aprendizado de máquina. Estes dados adicionais poderiam ser

- Verba utilizada
- Quantidade de pessoas em cada localidade
- Questões de repasses de verba entre governos federal, estaduais e municipais
- Empresas contratadas para realizar as obras

A utilização de modelos de machine learning diferentes como KNN e SVM tem potencial de obter resultados melhores, sendo necessário testá-los tanto com os dados utilizados neste trabalho, como com dados novos sugeridos neste item do documento.

5 Referências bibliográficas

BARBOZA, TEREZA NAZARÉ ULLOA; AGUIAR, CRISTIANE PEREIRA DE**. Inteligência Artificial na construção civil: benefícios, desafios e a importância do profissional na tomada de decisão. Caderno Pedagógico**, Curitiba, v. 22, n. 6, e15259, 2025. Disponível em: ri.ifam.edu.br › items › aa4a3cd2-e638-433d-91a3 . Acesso em: 27 out. 2025.

BRASIL. Lei nº 14.026, de 15 de julho de 2020. Atualiza o marco legal do saneamento básico e altera a Lei nº 9.984, de 17 de julho de 2000, para atribuir à Agência Nacional de Águas e Saneamento Básico (ANA) competência para editar normas de referência sobre o serviço de saneamento; a Lei nº 10.768, de 19 de novembro de 2003, para alterar o nome e as atribuições da [1] Agência Nacional de Águas (ANA); a Lei nº 11.445, de 5 de janeiro de 2007, para aprimorar as condições estruturais do saneamento básico no País; a Lei nº 12.305, de 2 de agosto de 2010, para tratar dos prazos para a disposição final ambientalmente adequada dos rejeitos; a Lei nº 13.089, de 12 de janeiro de 2015 (Estatuto da Metrópole), [2] para alterar a redação do parágrafo único do art. 8º; e a Lei nº 13.303, de 30 de junho de 2016, para alterar a redação do inciso I do caput do art. 24. **Diário Oficial da União**, Brasília, DF, 16 de julho de 2020. Seção 1, p. 1-14. Disponível em: https://www.jusbrasil.com.br/diarios/DOU/2020/07/16 . Acesso em: 31/03/2025.

FRIEDMAN, JEROME; HASTIE, TREVOR; TIBSHIRANI, ROBERT. ***The elements of statistical learning: data mining, inference, and prediction*.** 2. ed. Stanford, CA: Springer, 2009.

GALLARDO, SILVIA M. A. GUEDES. *Algumas das principais causas dos atrasos na entrega de obras públicas*. SÃO PAULO (Estado). Tribunal de Contas. Algumas das principais causas dos atrasos na entrega de obras públicas. In: **Revista do TCESP**, n. 142, p. 62. Out. 2018. Disponível em: https://www.tce.sp.gov.br/sites/default/files/publicacoes/Revista%20TCESP%20142%20-%20Outubro_2018.pdf. Acesso em: 27 out. 2025.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA**. ****Produto Interno Bruto - PIB****. [S. l.], 2024. Disponível em: https://www.ibge.gov.br/explica/pib.php . Acesso em: 31/03/2024.

Instituto Trata Brasil (ITB) (2024). **Coleta de esgoto sobe apenas 0,2 ponto percentual no país. Veja os melhores e os piores municípios destacados pelo Ranking do**

Saneamento 2024 https://tratabrasil.org.br/wp-content/uploads/2024/04/Release-Ranking-do-Saneamento-de-2024-TRATA-BRASIL-GO-ASSOCIADOS-V2.pdf

JAMES, G. ET AL. **An Introduction to Statistical Learning: with Applications in R.** New York: Springer, 2013.

MARSLAND, STEPHEN. **.** ***Machine learning, an algorithmic perspective.*** **(2nd ed.).** CRCPress – Taylor & Francis Group,2015

MITCHELL, TOM M**.** ***Machine learning*****.** [S. l.]: McGraw-Hill Education,1997

MORETTIN, P. A.; BUSSAB, W. O. **Estatística Básica.** 10. ed. São Paulo: Saraiva Uni, 2023.

NIETZSCHE, F. W. V. **Humano, demasiado humano**: Um Livro para Espíritos Livres. 2. ed. São Paulo: Companhia das Letras, 2017.

PROJECT MANAGEMENT INSTITUTE (PMI). **Guia do Conhecimento em Gerenciamento de Projetos (Guia PMBOK).** 7. ed. Estados Unidos: PMI, 2021.

RUSSEL, S., & NORVIG, P. . ***Artificial intelligence: A modern approach*** **(4th ed.).** Pearson Education, 2020

SILVA, JOÃO DA; SANTOS, MARIA DOS. Ferramentas de IA aplicadas à gestão de projetos e processos: uma experiência com chatbot. **Revista Brasileira de Gerenciamento de Projetos**, Rio de Janeiro, v. 10, n. 2, p. 45-60, 2024. Disponível em: https://www.collinsdictionary.com/dictionary/portuguese-english/artigo. Acesso em: 27 out. 2025.

TEIXEIRA, J. C., OLIVEIRA, G. S. DE,VIALI, A. DE M., & MUNIZ, S. S. . **Estudo do impacto das deficiências de saneamento básico sobre a saúde pública no Brasil no período [1] de 2001 a 2009** / Study of the impact of deficiencies of sanitation on public health in Brazil from 2001 to 2009. [2] *Engenharia Sanitaria e Ambiental*, *19*(01), 1-95.2014