



UNIVERSIDADE ESTADUAL PAULISTA

"JÚLIO DE MESQUITA FILHO"

Faculdade de Engenharia e Ciências de Guaratinguetá

MATHEUS CERQUEIRA DE JESUS

**Machine Learning para estimar o número de internações por doenças respiratórias em
Cuiabá-MT**

Guaratinguetá

2023

Matheus Cerqueira de Jesus

**Machine Learning para estimar o número de internações por doenças respiratórias em
Cuiabá-MT**

Trabalho de Graduação apresentado ao Conselho de Curso de Graduação em Engenharia Elétrica da Faculdade de Engenharia e Ciências do Campus de Guaratinguetá, Universidade Estadual Paulista, como parte dos requisitos para obtenção do diploma de Graduação em Engenharia Elétrica .

Orientador: Profº Dr. Paloma Maria Silva Rocha
Rizol

Guaratinguetá
2023

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
CAMPUS DE GUARATINGUETÁ

MATHEUS CERQUEIRA DE JESUS

ESTE TRABALHO DE GRADUAÇÃO FOI JULGADO ADEQUADO COMO PARTE DO
REQUISITO PARA A OBTENÇÃO DO DIPLOMA DE "**GRADUANDO EM ENGENHARIA
ELÉTRICA** "

APROVADO EM SUA FORMA FINAL PELO CONSELHO DE CURSO DE GRADUAÇÃO EM
ENGENHARIA ELÉTRICA

Profº Dr. DANIEL JULIEN BARROS DA SILVA SAMPAIO
Coordenador

BANCA EXAMINADORA:

Profº Dr. Paloma Maria Silva Rocha Rizol
Orientador/UNESP-FEG

Eng. Taynara de Oliveira Castellões
UNESP-FEG

Profº Dr. Leonardo Mesquita
UNESP-FEG

Fevereiro , 2023

DADOS CURRICULARES

MATHEUS CERQUEIRA DE JESUS

NASCIMENTO 09/11/1999 - São José dos Campos / SP

FILIAÇÃO Antonio Sergio de Jesus
Maria do Socorro Cerqueira de Jesus

Dedico este trabalho à todos aqueles que assim como eu, acreditam que a educação é a ferramenta
mais poderosa na luta contra a opressão.

AGRADECIMENTOS

Eu agradeço, primeiramente, aos meus pais que sempre acreditaram em mim, provendo todo o suporte necessário para que eu alcance os meus objetivos.

Agradeço a minha orientadora Paloma Maria Silva Rocha Rizol e coorientadora Taynara de Oliveira Castellões por todo o apoio durante o desenvolvimento desse trabalho.

Agradeço aos professores e amigos Luis Rogerio de Oliveira Hein e Maurício de Oliveira Filho por todas as conversas e horas de trabalho juntos, vocês participaram da minha formação como pessoa.

Agradeço aos meus queridos amigos Tales Hiro Cardoso Ishida, Thales Vieira e Silva Lobo De Almeida, Vinicius Mancini e Matheus Vinicius Resende Nascimento, por todas as horas de estudo juntos e todos os bons momentos que tivemos.

Agradeço a todos os professores e funcionários da universidade que contribuíram para a minha formação.

Finalmente, agradeço a minha companheira Deborah Eberle dos Santos por todo o incentivo, paciência, momentos felizes e por me fazer ser uma pessoa melhor.

“Pouco conhecimento faz com que as pessoas se sintam orgulhosas. Muito conhecimento, com que se sintam humildes.”
(Leonardo da Vinci)

RESUMO

Uma estimativa realizada em 2021 aponta que a cada ano, a exposição à poluição atmosférica seja responsável por 7 milhões de mortes prematuras e milhões de anos de vida reduzidos de pessoas com uma vida saudável em todo o mundo. Outro agravante relacionado a poluição é o aumento da temperatura média do planeta que em 2021 foi 1,1°C acima da linha base pré-industrial. Devido as queimadas no ano de 2017 a cidade de Cuiabá-MT, Brasil enfrentou altas taxas de material particulado, valores muito maiores do que o mínimo considerado tolerável para os seres humanos. Esses altos valores contribuem para o surgimento de problemas respiratórios graves, como o broncoespasmo, pneumonia e bronquites. Em consideração a esse contexto, as demandas dos serviços de saúde estão cada vez maiores e situações em que os serviços operam acima da capacidade máxima estão cada vez mais frequentes. Os sistemas e serviços de saúde necessitam de recursos essenciais para atuar, incluindo informações hospitalares, doenças e previsões. Como solução a esse problema, foram analisados modelos de aprendizado de máquina e uma rede neural artificial, comparando os resultados obtidos pelo modelo de Regressão Linear, Árvore de Decisão, Floresta Aleatória e uma LSTM. O modelo que obteve o melhor desempenho foi a LSTM utilizando uma abordagem de séries temporais com um RMSE de aproximadamente 3 internações.

PALAVRAS-CHAVE: Doenças Respiratórias. Poluentes Atmosféricos. Cuiabá-MT. Aprendizado de Máquina. Rede Neural Artificial.

ABSTRACT

An estimate made in 2021 indicates that each year, exposure to air pollution is responsible for 7 million premature deaths and millions of reduced life years of people with healthy living around the world. Another aggravating factor related to pollution is the increase in temperature average of the planet which in 2021 was 1.1°C above the pre-industrial baseline. Due to fires in 2017, the city of Cuiabá faced high rates of particulate matter, much higher values than the minimum considered tolerable for humans. These high values contribute to the emergence of serious respiratory problems, such as bronchospasm, pneumonia and bronchitis. Given this context, the demand for health services is increasing and situations in which services operating above maximum capacity are increasingly frequent. The systems and health care services need essential resources to act, including hospital information, diseases and predictions. As a solution to this problem, machine learning models were analyzed and a neural network, comparing the results obtained by the Linear Regression, Decision Tree, Random Forest and an LSTM. The model that obtained the best performance was the LSTM using a time series approach with an RMSE of approximately 3 admissions.

KEYWORDS: Respiratory Diseases. Atmospheric Pollutants. Cuiabá-MT Machine Learning. Artificial Neural Network.

LISTA DE ILUSTRAÇÕES

Figura 1	Gráfico da quantidade de artigos publicados	15
Figura 2	Distribuição de artigos publicados por país	16
Figura 3	Estrutura das árvores de decisão	19
Figura 4	Estrutura da <i>random florest</i>	20
Figura 5	Exemplo de um conjunto de treino e teste	21
Figura 6	Exemplo do funcionamento da validação cruzada	22
Figura 7	Exemplo de uma unidade LSTM	23
Figura 8	Interface do SISAM	26
Figura 9	Dados com 4 medições diárias	27
Figura 10	Dados filtrados e tratados	27
Figura 11	Interface de coleta de dados do Datasus	28
Figura 12	Arquivos para <i>download</i> do Datasus.	28
Figura 13	Conversão dos arquivos em .dbf.	29
Figura 14	Conversão dos arquivos em .csv.	29
Figura 15	Tratamento final dos dados de doenças respiratórias.	30
Figura 16	Tabela com todas as <i>features</i> necessárias para os treinamentos dos modelos	30
Figura 17	Comparação entre estratificar o <i>dataset</i> por ano e apenas randomizar	30
Figura 18	Construção do modelo de regressão linear	31
Figura 19	Construção do modelo de árvore de decisão	32
Figura 20	Construção do modelo de floresta aleatória	32
Figura 21	Função de <i>tuning</i> para construção do modelo de LSTM	33
Figura 22	Diagrama de dispersão dos dados reais e dos previstos do modelo LR	35
Figura 23	Diagrama de dispersão dos dados reais e dos previstos do modelo DT	36
Figura 24	Diagrama de dispersão dos dados reais e dos previstos do modelo RF	37
Figura 25	Diagrama de dispersão dos dados reais e dos previstos do modelo LSTM	38

LISTA DE TABELAS

Tabela 1 – Publicações no Scopus	15
Tabela 2 – Aplicações de IAs na medicina	19
Tabela 3 – Informações da cidade de Cuiabá	25
Tabela 4 – Correlação linear entre as <i>features</i> e a variável de saída	31
Tabela 5 – Transformação dos dados para aprendizado supervisionado (lag=3 dias)	34
Tabela 6 – Métricas do modelo de regressão linear	35
Tabela 7 – Métricas do modelo de árvore de decisão	36
Tabela 8 – Métricas do modelo de floresta aleatória	36
Tabela 9 – Métricas do modelo LSTM	37
Tabela 10 – Resultados dos modelos	37

LISTA DE ABREVIATURAS E SIGLAS

TCC	Trabalho de Conclusão de Curso
UNESP	Universidade Estadual Paulista
IA	Inteligência Artificial
SISAM	Sistema de Informações Ambientais Integrado a Saúde
DATASUS	Departamento de Informática do Sistema Único de Saúde
LR	<i>Linear Regression</i>
DT	<i>Decision Tree</i>
RF	<i>Random Forest</i>
LSTM	<i>Long Short Term Memories</i>
ANN	<i>Artificial Neural Network</i>
SUS	Sistema Único de Saúde
SIH	Sistema de Informações Hospitalares
RNN	<i>Recurrent Neural Network</i>
MAE	<i>Mean Absolute Error</i>
MAPE	<i>Mean Absolute Percentage Error</i>
MSE	<i>Mean Squared Error</i>
RMSE	<i>Root Mean Squared Error</i>
RD	Reduzida
AIH	Autorização de Internação Hospitalar
CID	Classificação Internacional de Doenças
API	<i>Application Programming Interface</i>

LISTA DE SÍMBOLOS

R\$	Unidade Monetária Brasileira (Real)
°C	Graus Celsius
m	Unidade de medida de comprimento: Metro
mm	Unidade de medida de comprimento: Milímetros
g	Unidade de medida de massa: Grama
ppb	Unidade de medida de concentração: Partes por bilhão
µg	Unidade de medida de massa: Micrograma
m ³	Unidade de medida de volume: Metro Cúbico
km ²	Unidade de medida de área: Quilómetro quadrado

SUMÁRIO

1	INTRODUÇÃO	14
1.1	OBJETIVOS	14
1.2	JUSTIFICATIVAS	15
1.3	DELIMITAÇÕES DA PESQUISA	16
1.4	ESTRUTURA DO TRABALHO	16
2	FUNDAMENTAÇÃO TEÓRICA	18
2.1	PYTHON E CIÊNCIA DE DADOS	18
2.2	INTELIGÊNCIA ARTIFICIAL NO CONTEXTO DA SAÚDE	18
2.2.1	<i>Decision Trees</i>	19
2.2.2	<i>Random Forest</i>	20
2.2.3	<i>Cross Validation</i>	21
2.3	PREVISÃO DE SÉRIES TEMPORAIS NO CONTEXTO DA SAÚDE	22
2.3.1	<i>Long Short Term Memories</i>	23
2.4	MÉTRICAS PARA AVALIAR OS MODELOS DE REGRESSÃO	23
2.4.1	Erro quadrático médio	24
2.4.2	Raiz do erro quadrático médio	24
2.4.3	Erro médio absoluto	24
3	MATERIAIS E MÉTODOS	25
3.1	DESCRIÇÃO DO PROBLEMA	25
3.2	OBTENÇÃO DOS DADOS	26
3.3	MODELAGEM DO PROBLEMA	28
3.3.1	Primeira abordagem	29
3.3.1.1	Modelo de regressão linear	31
3.3.1.2	<i>Decision Tree</i>	31
3.3.1.3	<i>Random Forest</i>	32
3.3.2	Segunda abordagem	32
3.3.2.1	LSTM	33
4	ANÁLISE DOS RESULTADOS OBTIDOS	35
5	CONCLUSÃO	39
5.1	PROPOSTAS PARA FUTURAS PESQUISAS	39
	REFERÊNCIAS	40

1 INTRODUÇÃO

Os últimos séculos foram marcados por períodos de intensos desenvolvimentos tecnológicos, sendo impulsionados pelas revoluções industriais. As indústrias são responsáveis por produzir a maior parte dos produtos essenciais para os seres humanos, bem como o desenvolvimento de novos produtos. Para a produção de carros cada vez mais potentes, escalas enormes de alimentos, dispositivos digitais, produtos farmacêuticos é necessária a produção de energia para o funcionamento das máquinas das indústrias, automóveis para o transporte de cargas e pessoas, e atender as populações das cidades e do campo.

No entanto, de acordo com a (AGENCY, 2023) a matriz energética mundial é majoritariamente composta por fontes de energia não renováveis e altamente poluentes, como os combustíveis fósseis, sendo eles, carvão mineral, petróleo e gás natural. Como consequência do consumo constante dessas matrizes energéticas os níveis de poluentes na atmosfera passaram a aumentar. E com isso a necessidade de realizar estudos para entender a influência desses poluentes na saúde humana.

Segundo a (UNIDAS, 2021) uma estimativa realizada em 2021 aponta que a cada ano, a exposição à poluição atmosférica seja responsável por 7 milhões de mortes prematuras e milhões de anos de vida reduzidos de pessoas com uma vida saudável em todo o mundo. Outro agravante relacionado a poluição é o aumento temperatura média do planeta que segundo a (UNIDAS, 2022) em 2021 a temperatura média do planeta foi de 1,1°C acima da linha base pré-industrial. Diante dessa situação, (XU; HU; TONG, 2014) apontam que as altas e baixas temperaturas estão associadas com o aumento da incidência de casos de pneumonia e que as altas variações de temperatura diurnas e entre dias podem afetar o funcionamento do sistema respiratório.

À vista desse contexto, a demanda dos serviços de saúde estão cada vez maiores e situações em que os serviços operam acima da capacidade máxima estão cada vez mais frequentes. Conforme apresentado por (WHO, 2010) os sistemas e serviços de saúde necessitam de recursos essenciais para atuar, incluindo informações hospitalares, doenças e previsões. No entanto, de acordo com (SOYIRI; REIDPATH; SARRAN, 2012), os hospitais, serviços de saúde e fornecedores geralmente não estão adequadamente informados quando entram em uma situação de demanda acima do normal.

Este trabalho, reuniu dados e informações diárias relacionados a qualidade do ar, temperatura ambiente e internações ocorridas em hospitais de Cuiabá no estado do Mato Grosso entre 01/01/2013 e 31/12/2018 para treinar modelos de *Machine Learning* e redes neurais afim de prever o número de internações por doenças respiratórias nos dias subsequentes a uma determinada data, utilizando como dados de entradas os valores de qualidade do ar e temperatura. Os resultados dos diferentes modelos foram comparados, observando principalmente a raiz do erro quadrático médio(RMSE).

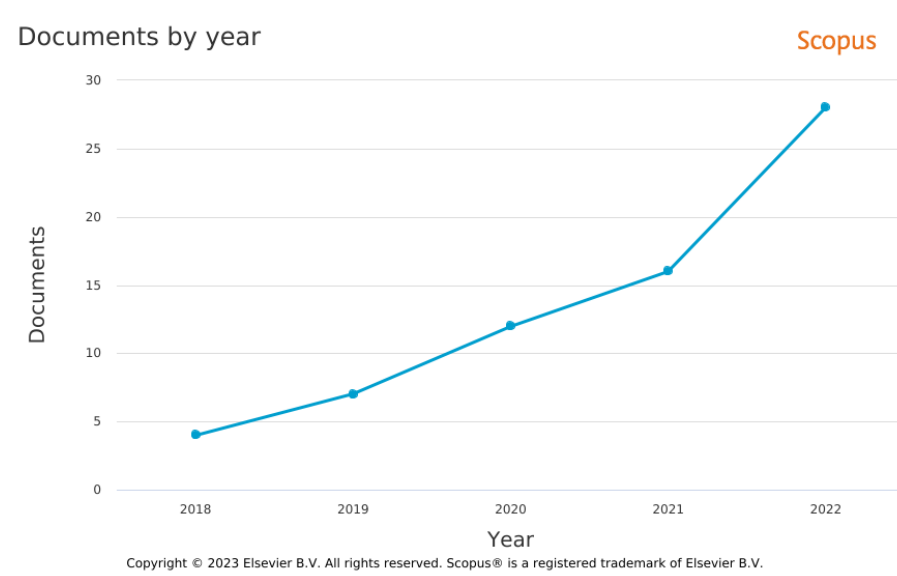
1.1 OBJETIVOS

Este trabalho tem como objetivo utilizar técnicas de *Machine Learning* e redes neurais artificiais a fim de prever internações hospitalares de pacientes com doenças respiratórias causadas por poluições atmosféricas.

1.2 JUSTIFICATIVAS

Na plataforma Scopus foram realizadas três pesquisas configurando a busca para título do artigo, resumo e palavras-chave, a fim de verificar o contexto de trabalhos realizados na academia com temas semelhantes aos deste trabalho, conforme apresentado na Tabela 1. As pesquisas abrangeram o período dos anos de 2012 até 2022 utilizando as palavras chave: Machine Learning, Respiratory Diseases e Air Quality. Nota-se que à medida em que as palavras chaves tornam a busca na plataforma mais específica a quantidade de artigos publicados diminui consideravelmente. Para a pesquisa com as palavras chave: *Machine Learning*, *respiratory diseases* e *air quality*, foram publicados somente 67 artigos, isso mostra ser um tema que ainda não foi muito pesquisado, apesar de estar em crescimento, conforme apresentado na Figura 1.

Figura 1 – Gráfico da quantidade de artigos publicados



fonte: Scopus (2023)

Ao adicionar Cuiabá nas palavras chave da busca, não foi retornado nenhum resultado, isso indica que o trabalho realizado é inédito.

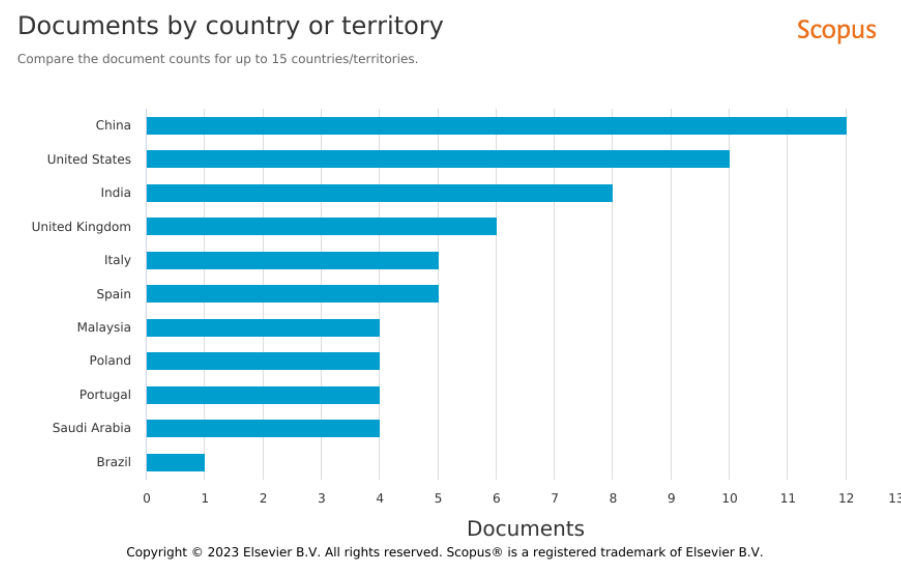
Tabela 1 – Publicações no Scopus

Pesquisa	Período	Total de publicações
"Machine Learning"	2012 - 2022	507.100
"Machine Learning"e "Respiratory Diseases"	2012 - 2022	2.338
"Machine Learning"e "Respiratory Diseases"e "Air Quality"	2012 - 2022	67

fonte: Scopus (2023)

A Figura 2 apresenta a distribuição de trabalhos acadêmicos por países, analisando-a verifica-se que o Brasil tem apenas um trabalho publicado na plataforma, o que aponta para uma grande necessidade por novas pesquisas.

Figura 2 – Distribuição de artigos publicados por país



fonte: Scopus (2023)

1.3 DELIMITAÇÕES DA PESQUISA

Ao decorrer do trabalho foram verificadas algumas limitações. Os dados utilizados para o treinamento dos modelos de *Machine Learning* e (ANN) possuem colunas com informações sobre a qualidade do ar, temperatura, umidade e internações por doenças respiratórias. Os dados de qualidade do ar são estimativas realizadas pelo SISAM para a cidade de Cuiabá-MT, assim considerou-se que todos os habitantes estão submetidos as mesmas condições.

Ao considerar que todas as pessoas estão influenciadas pelas mesmas condições, desconsidera-se as informações de onde essas pessoas vivem e com o que elas trabalham. Por exemplo, uma pessoa que trabalhe em uma marcenaria está sobre condições respiratórias piores do que outras pessoas e essa influência não foi levada em conta nos modelos por não possuir dados.

Os dados de internações representam apenas as ocorridas nos hospitais da rede pública (SUS). Dessa forma, todas as internações por doenças respiratórias ocorridas na rede particular de hospitais não foram contabilizadas. Ainda, as informações das internações por doenças respiratórias são diagnosticadas pelos médicos e inseridas no sistema SIH/SUS, portanto pode haver algum engano ao ser inserida uma informação manualmente no sistema.

1.4 ESTRUTURA DO TRABALHO

Este trabalho está dividido em 5 partes:

- Introdução
- Fundamentação teórica

Neste capítulo, será abordado os principais conceitos utilizados para a confecção do trabalho, comentando sobre modelos de aprendizado de máquina, redes neurais artificiais e métricas para avaliações dos modelos.

- Materiais e métodos

Aqui será redigida a descrição do problema, como os dados foram coletados, tratados e utilizados nos modelos.

- Análise dos resultados obtidos

Comparação dos resultados de cada modelo analisando as métricas utilizadas e escolhendo o modelo que obteve os melhores resultados

- Conclusão

Neste último capítulo será feita uma conclusão do trabalho com propostas de pesquisas futuras.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo foi escrita uma descrição resumida das ferramentas, técnicas e conceitos utilizados no desenvolvimento do trabalho.

2.1 PYTHON E CIÊNCIA DE DADOS

Python é uma linguagem de programação de computadores multiparadigma e de código aberto (*open source*), que de acordo com (LUTZ, 2013) é otimizada para programar de forma produtiva, ler e entender os códigos com facilidade e qualidade de *software*.

Python é a linguagem mais utilizada do mundo segundo o rank da linguagens de (CARBONNELLE, 2023) ganhando de Java, Java Script e C++. Essa liderança explica a extensa comunidade que a linguagem possui, sendo isso uma vantagem para quem utiliza, uma vez que é possível encontrar milhares de exemplos de código para uma possível aplicação que alguém precise.

Uma ótima vantagem da linguagem é a incrível quantidade de bibliotecas disponibilizadas pela comunidade para as mais variadas aplicações. Isso torna o Python uma ótima ferramenta para trabalhar com inteligência artificial, *Machine Learning* e *deep learning*. Algumas das principais bibliotecas e *frameworks* utilizadas são Pandas e Numpy para a manipulação de dados, TensorFlow e Scikit Learning para a construção de modelos de *Machine Learning*, como os que serão abordados nos capítulos 2.2 e 2.3.1.

2.2 INTELIGÊNCIA ARTIFICIAL NO CONTEXTO DA SAÚDE

Inteligência artificial tem sido um assunto amplamente abordado pelas pessoas em vídeos, notícias e filmes. Muitos acreditam que as IAs podem ser a solução para todos os problemas existentes na sociedade. No entanto, é importante saber o que realmente é inteligência artificial e para o que realmente pode ser usada.

O professor (XIAO, 2022) define em seu livro inteligência artificial como sendo uma área de estudo dentro ciências da computação que possui o objetivo de fazer com que máquinas aprendam a interpretar e resolver problemas de maneira similar ao ser humano. Da mesma forma que um ser humano, aprende ao tentar resolver problemas e obter novas informações, uma inteligência artificial deve tomar uma ação a medida em que recebe novas informações e aprende a fim de melhorar sua performance.

Alguns exemplos de inteligência artificial presente no dia a dia de muitas pessoas são as assistentes virtuais como a Alexa da Amazon e a Cortana da Microsoft, outro exemplo são as ferramentas de anúncio na internet que aprendem com informações sobre os usuários e enviam anúncios de maior interesse. Mas como as IAs, estão relacionadas com a saúde das pessoas e os serviços de assistência médica?

De acordo com Trishan Panch:

A Inteligência Artificial e o aprendizado de máquina têm o potencial de ser o catalisador da transformação dos sistemas de saúde para melhorar a eficiência e a eficácia, criar margem para a cobertura universal de saúde e melhorar os resultados. (PANCH; SZOLOVITS; ATUN, 2018, p.1)

Nos sistemas de saúde existem processos que podem utilizar IAs. Dois exemplos são segundo (PANCH; SZOLOVITS; ATUN, 2018) o diagnóstico de doenças dos pacientes, realizando uma tarefa de classificação, se está ou não com a doença. O outro processo se dá durante o tratamento, envolvendo predição de uma melhora ou piora no quadro do paciente, monitorando os dados vitais. Algumas das aplicações existentes foram apresentadas na tabela 2.

Tabela 2 – Aplicações de IAs na medicina

Diagnóstico	Prognóstico e predição
Análise de imagem: Mamografia	Hospitalização por doença cardíaca
Análise de sinais: Monitoramento intraparto	Risco de acidente cardiovascular
Análise de imagem: identificação da retinopatia diabética	Predição de resultados em câncer colorretal

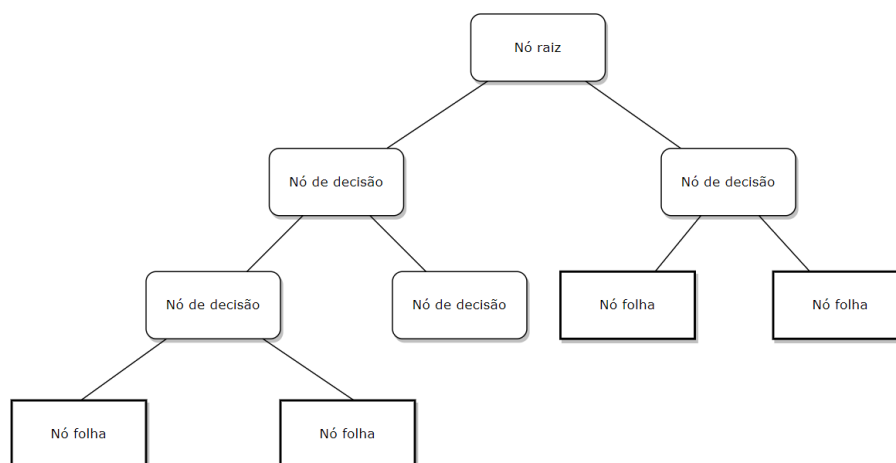
fonte: Adaptado de (PANCH; SZOLOVITS; ATUN, 2018)

Diante disso, é notável que as IAs podem contribuir para melhorar os serviços de saúde, impactando a vida de milhares de pessoas. A seguir serão abordadas algumas técnicas de aprendizado de máquinas utilizadas neste trabalho.

2.2.1 Decision Trees

Árvores de decisão (*Decision Trees*) são algoritmos de *machine learning* muito versáteis, que podem ser utilizados tanto para resolver problemas de classificação quanto de regressão. Como o próprio nome diz as *Decision Trees* possuem uma estrutura hierarquia em árvore, contendo o nó raiz (*root node*), as ramificações, nós de decisão (*decision nodes*) e as folhas (*leaf nodes*). Uma estrutura simples desse algoritmo está apresentada na Figura 3.

Figura 3 – Estrutura das árvores de decisão



fonte: Produção do próprio autor.

O modelo de *Decision Tree* utilizado nesse trabalho realizou uma tarefa de regressão, ou seja, uma *Regression Tree*, nessa variação todos os nós da árvore possuem valores numéricos. Imaginando um conjunto de pontos (x, y) , o primeiro passo para montar a árvore é obter o nó raiz, sendo encontrado ao calcular iterativamente para o valor médio entre dois x adjacentes o erro quadrático médio (MSE) com a média dos valores de y a esquerda do ponto médio e a média dos valores de y a direita do ponto médio. O valor médio que obtiver o menor MSE será o nó raiz da árvore.

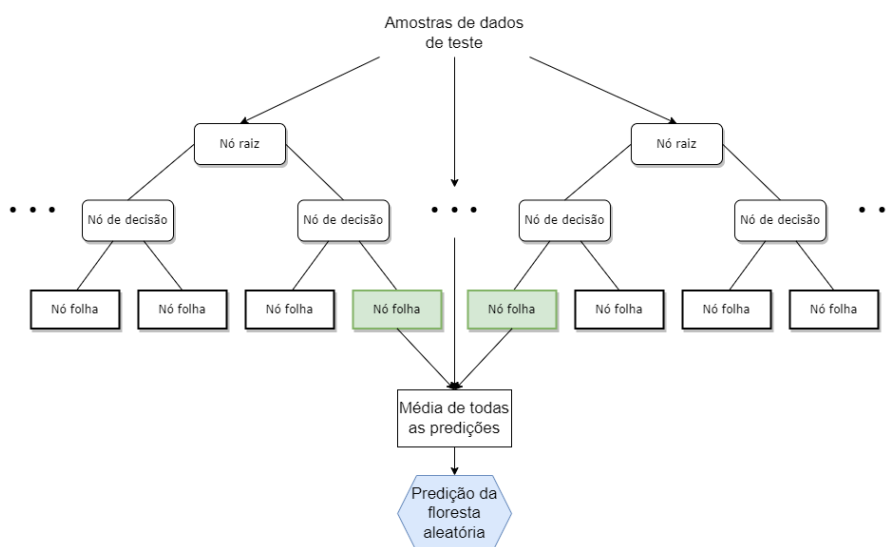
Para finalizar todo o processo de construção o processo é repetido para o lado esquerdo do nó raiz e para o lado direito até que não sobrem mais valores possíveis para as folhas.

2.2.2 Random Forest

O *random forest* (floresta aleatória) é um algoritmo de machine learning do tipo aprendizagem supervisionada que pode ser utilizado tanto para resolver problemas de classificação quanto de regressão. O RF foi desenvolvido, com o intuito de resolver algumas desvantagens observadas nos modelos de *decision tree*.

O conceito de florestas está na construção da *random forest*, em que é construída realizando um agrupamento de *decision tree* durante o treinamento, chamado de método *ensemble*, ou seja, reúne a predição dos múltiplos algoritmos pela média, resultando em uma predição mais acurada do que apenas uma *decision tree*. Conforme apresentado na Figura 4:

Figura 4 – Estrutura da *random forest*



fonte: Produção do próprio autor.

O funcionamento do algoritmo ocorre da seguinte forma, primeiramente, deve-se escolher o número de árvores de decisão, então selecionam-se os dados de forma aleatória podendo haver repetições, repetindo o processo para o número de *decision tree* escolhido. Após isso, as árvores são construídas a partir dos dados aleatórios selecionados para cada uma, por isso nomeia-se floresta aleatória.

2.2.3 Cross Validation

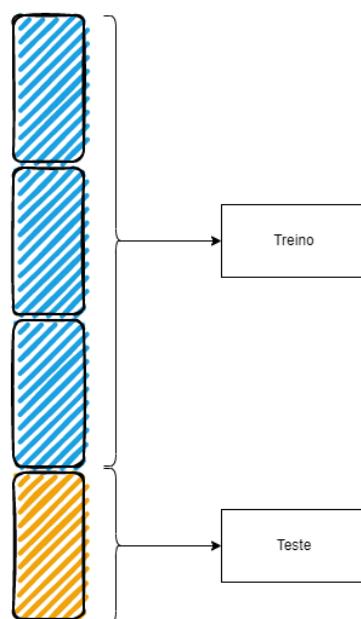
Existem diferentes modelos de *machine learning* disponíveis que podem ser empregados na resolução de problemas, e como consequência dessas possibilidades aparecem algumas dúvidas:

- Qual modelo é melhor?
- Qual modelo terá o melhor desempenho?
- Qual será o mais estável ao receber *inputs* inéditos?

Antes de realizar o treino e o teste do modelo é necessário preparar o conjunto de dados em que filtrar ruídos e tratar valores nulos são exemplos de passos que podem ser aplicados. Após essa etapa, o conjunto de dados é randomizado e dividido em teste e treino, de acordo com (GÉRON, 2022) é comum dividir o conjunto de dados em 80% treino e 20% teste, no entanto em casos que a quantidade de dados é massiva, diminuir o percentual dos dados de teste pode ser uma boa prática.

Ao realizar esses passos será obtido um conjunto de acordo com a Figura 5.

Figura 5 – Exemplo de um conjunto de treino e teste



fonte: Produção do próprio autor.

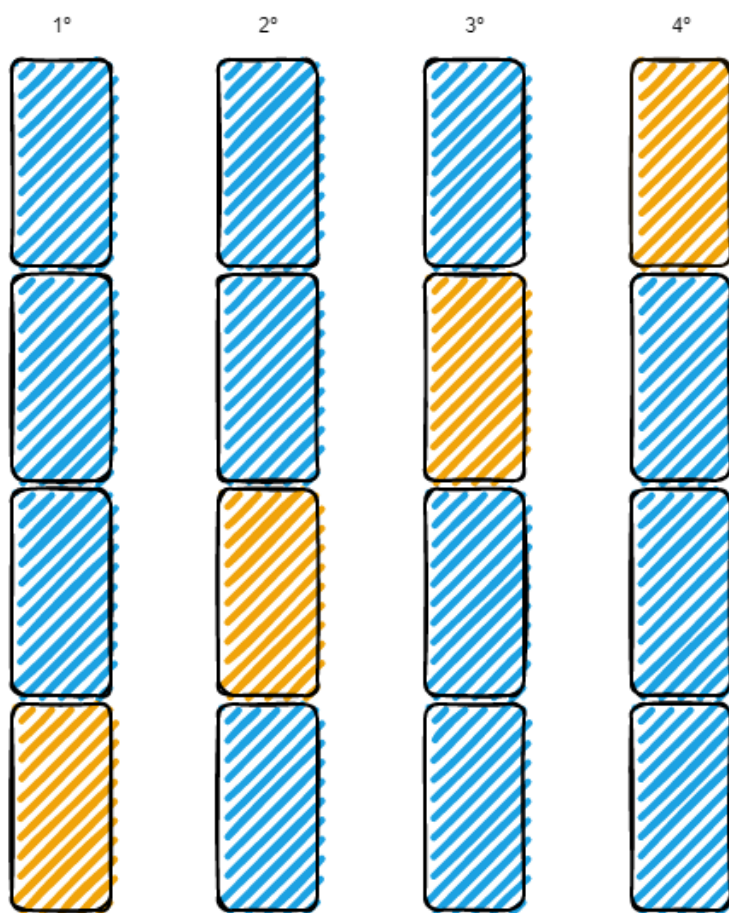
Um possível problema dessa abordagem é a utilização de apenas a última parte do *dataset* para o treino, isso porque, mesmo que apresente um erro de generalização pequeno, ao colocar o modelo em produção dados completamente inéditos e diferentes do conjunto de teste podem aparecer, mostrando que o modelo não generaliza tão bem quanto o esperado.

Uma solução para esse problema é a validação cruzada (*cross validation*), uma técnica muito empregada para avaliar o desempenho de modelos com relação a todo o conjunto de dados, a fim de verificar qual modelo obtém a melhor generalização.

A validação cruzada consiste em particionar o conjunto de dados em partes iguais, sendo a quantidade de partes escolhidas arbitrariamente. A Figura 6 exemplifica uma *4-fold cross validation*,

ou seja, o conjunto de dados dividido em 4 partes. Os modelos são então treinados 4 vezes, variando a ordem das partições de treino e teste, ao final faz-se a média do erro de generalização e verifica qual modelo mostrou ter o melhor desempenho em todo o conjunto de dados. Esse será o melhor a ser utilizado.

Figura 6 – Exemplo do funcionamento da validação cruzada



fonte: Produção do próprio autor.

2.3 PREVISÃO DE SÉRIES TEMPORAIS NO CONTEXTO DA SAÚDE

Uma série temporal, ou série histórica, é definida segundo (LATORRE; CARDOSO, 2001) como uma sequência de dados obtidos em intervalos regulares durante um determinado período. A série pode ser obtida tanto ao realizar medições por sensores, por exemplo, temperatura, pressão, tanto por contagens, por exemplo, o número de internações mensais por doenças respiratórias.

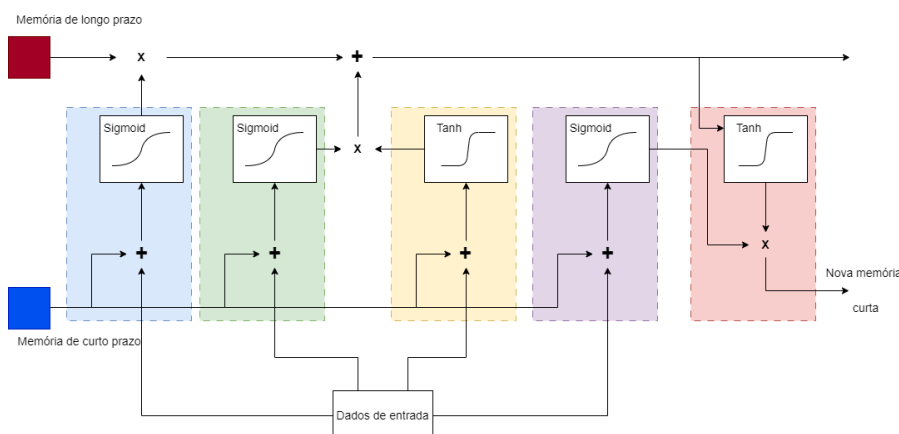
A previsão de séries temporais é utilizada no âmbito dos sistemas de saúde em diferentes contextos, como previsão de epidemias de doenças antigas e novas e previsão das demandas no sistema de saúde, por internações, atendimentos e cirurgias. Conforme (SOYIRI; REIDPATH; SARRAN, 2012) os serviços de saúde geralmente estão mal informados e com recursos insuficientes para se adaptar a períodos como alta demanda, por isso, utilizar uma abordagem de séries temporais, pode ser útil para promover melhores informações e ajudar em tomada de decisões mais eficientes por médicos e administradores do sistema de serviço de saúde.

2.3.1 Long Short Term Memories

As *Long Short Term Memories* são um tipo específico de rede neural recorrente (RNN) que tem como objetivo resolver o problema do *explode/vanishing gradient* da RNN original. Para isso elas utilizam dois caminhos diferentes para realizar uma predição, o caminho de memória curta *short term* e o caminho de memória longa *long term*.

As LSTMs possuem uma unidade mais complicada do que as RNNs tradicionais, conforme apresentado na Figura 7.

Figura 7 – Exemplo de uma unidade LSTM



fonte: Adaptado de (STARMER, 2022).

As memórias de curto e longo prazo podem ser iniciadas com um valor arbitrário, geralmente sendo 0. Então, analisando de forma sequencial, o primeiro bloco (azul) da unidade LSTM, é responsável por determinar quanto da memória de longo prazo será lembrada, isso porque a saída desse bloco é uma função sigmoideal, portanto quando o dado de entrada soma-se a memória de curto prazo, são aplicados na função sigmoideal resultando em um valor entre 0 e 1, esse primeiro bloco é chamado de *forget gate*. A próxima etapa de processamento da unidade são o conjunto de blocos verde e amarelo (*input gate*), sendo a sua função calcular quanto um valor a ser somado na memória de longo prazo, o bloco verde gera a porcentagem a ser somada e o bloco amarelo uma memória em potencial. A última etapa da LSTM é o *output gate* que relaciona a memória de curto prazo a entrada e a memória de longo prazo para produzir uma nova memória de curto prazo. A saída da LSTM *unit* pode ser utilizada tanto para a predição do modelo ou para entrada para a próxima unidade em uma rede neural artificial LSTM.

Resumindo, de acordo com (GÉRON, 2022) as unidades LSTM aprendem a reconhecer entradas importantes da série temporal, armazenando-as na memória de longo prazo, para preservar-lá por um certo tempo e assim, reutiliza-lá sempre que necessário. Essa estrutura robusta são o motivo da LSTM obter performances tão boas em detectar padrão de longo prazo em séries temporais, textos longos e gravações de áudio.

2.4 MÉTRICAS PARA AVALIAR OS MODELOS DE REGRESSÃO

Durante a estimativa dos parâmetros de um modelo de regressão, é necessária a utilização de métricas para a validação da sua performance. Nesse contexto, as métricas mais utilizadas são as

relacionadas ao erro entre a predição e o valor esperado, sendo algumas métricas, erro médio absoluto (MAE), percentual do erro médio absoluto (MAPE), erro quadático médio (MSE) e raiz do erro quadático médio (RMSE).

2.4.1 Erro quadrático médio

O MSE é calculado pela equação 1:

$$\sum_{i=1}^D (y_i - \hat{y}_i)^2 \quad (1)$$

Analisando a equação, percebe-se que como o erro está elevado ao quadrado nunca haverá um valor negativo e um outro efeito dessa métrica é punir o modelo quanto maior for o erro, uma vez que, um número grande elevado ao quadrado será um número ainda maior.

2.4.2 Raiz do erro quadrático médio

O RMSE é calculado pela equação 2:

$$\sqrt{\sum_{i=1}^D (y_i - \hat{y}_i)^2} \quad (2)$$

Essa métrica é mais empregada para analisar a performance do modelo e mostrar os resultados, enquanto o MSE é utilizado como *loss function*. Outro ponto importante é a unidade do erro que é a mesma do valor de saída.

2.4.3 Erro médio absoluto

O MAE é calculado pela equação 3:

$$\sum_{i=1}^D |x_i - y_i| \quad (3)$$

Diferentemente do MSE, essa métrica não intensifica o erro, quanto maior for, mas varia linearmente com o aumento ou diminuição do erro.

3 MATERIAIS E MÉTODOS

Nesse capítulo será realizada toda a descrição do problema, desde a escolha da cidade de Cuiabá, as queimadas ocorridas, até os principais poluentes analisados.

3.1 DESCRIÇÃO DO PROBLEMA

Cuiabá é a capital do estado do Mato Grosso, localizada na região centro-oeste do Brasil e reconhecida por ser o principal polo industrial, comercial e de serviços do seu estado. A Tabela 3 apresenta algumas informações da cidade:

Tabela 3 – Informações da cidade de Cuiabá

Métricas	Valores
Área territorial	5.077,181 km ² [2021]
População estimada	623.614 pessoas [2021]
Densidade demográfica	157,66 hab/km ² [2010]
Mortalidade infantil	12,92 óbitos por mil nascidos vivos [2020]
PIB per capita	42.918,31 R\$ [2020]

fonte: Adaptado de (IBGE, 2023a)

Segundo dados do (IBGE, 2023b) 80,2% dos domicílios possuem esgotamento sanitário adequado, apenas 39,6% dos domicílios urbanos tem arborização e 34,3% com urbanização adequada.

Ainda, de acordo com uma notícia de (GROSSO, 2017) o índice de material particulado durante queimadas no ano de 2017 oscilou entre 100 e 140 µg/m³, valor muito maior do que o mínimo considerado tolerável pelo ser humano de 25µg/m³. Esses valores altos observados contribuem para o surgimento de problemas respiratórios graves, como o broncoespasmo, que causa o estreitamento das vias respiratórias e, portanto, dificuldade de respirar. Além disso, essa poluição causa irritação das vias aéreas, tornando-as mais suscetíveis a vírus e bactérias levando muitas pessoas a ter infecções graves como a pneumonia. Sendo outros problemas comuns as crises de asma, bronquites, crises alérgicas, rinites, sinusites e irritação nos olhos.

Diante desse contexto, tornou-se evidente a importância do presente tema desse trabalho. Ao analisar como diferentes modelos de aprendizado de máquina e uma rede neural artificial performam com os dados utilizados e sua capacidade de predição de novas internações por doenças respiratórias, pode ser uma importante ferramenta para, principalmente, ajudar a salvar mais vidas, gerenciar recursos e melhorar a qualidade de vida da população.

Os principais poluentes utilizados nos treinamentos dos modelos foram o ozônio em partes por bilhão e o material particulado de 2,5 micrômetros (µg/m³), além dessas *features*, o banco de dados, também conta com dióxido de enxofre (SO₂ [µg/m³]) e monóxido de carbono (CO [ppb]). As informações climáticas e ambientais são a temperatura média diária [°C], umidade relativa percentual, precipitação pluviométrica diária [mm] e os focos de queimadas na cidade. Por fim, o banco contém a quantidade diária de internações ocorridas por doenças e complicações respiratórias nos hospitais públicos de Cuiabá.

3.2 OBTENÇÃO DOS DADOS

Os dados utilizados para o projeto foram retirados de dois bancos de dados governamentais: O Sistema de Informações Ambientais Integrado a Saúde (SISAM) e o Sistema de informações hospitalares do SUS (SIHSUS).

O site do SISAM disponibiliza na aba Dados/Downloads uma interface para realizar a filtragem dos dados desejados e fazer a requisição de download, conforme apresentado na Figura 8:

Figura 8 – Interface do SISAM

fonte: SISAM (2023)

No sistema é possível escolher um período de até 1 ano, o estado e a cidade, as variáveis meteorológicas e poluentes. Então, clica-se em "gerar arquivo csv" para realizar o download de todos os dados selecionados.

O projeto utilizou 6 anos de dados que englobou os períodos de primeiro de Janeiro de 2013 até 31 de dezembro de 2018, dessa forma fez-se 6 vezes o download dos dados para cada ano até 2018. Foi necessário juntar todos os arquivos em um único, isso foi feito por meio da biblioteca Pandas do Python.

Para cada dia haviam 4 horários de medições, por isso, o tratamento escolhido foi calcular a média diária para todas as colunas dos dados. A Figura 9, apresenta os dados antes do tratamento com 8.764 linhas e possuindo valores nulos nas colunas precipitacao_mmdia e focos_queimada.

A Figura 10 apresenta os dados agrupados por dia, sendo cada valor diário a média das 4 medições diárias, os dados nulos da coluna de focos de queimadas foram substituídos por 0. Dessa forma, o conjunto de dados climáticos e de poluentes ficou com 2.191 linhas que são exatamente 6 anos.

A obtenção dos dados de internações por doenças respiratórias foi através do sistema de transferência de arquivos do site Datasus. No site é necessário selecionar a fonte: Sistema de informações hospitalares do SUS (SIHSUS), modalidade: Dados, tipo de arquivo: RD - AIH Reduzida, o período desejado, nesse caso 2013 até 2018, quais meses do ano, sendo para esse projeto todos os meses e o estado: MT (Mato Grosso), conforme apresentado na Figura 11.

Figura 9 – Dados com 4 medições diárias

```
df_air.drop(columns=columns_to_drop, inplace=True)
df_air
```

	datahora	co_ppb	o3_ppb	pm25_ugm3	so2_ugm3	precipitacao_mmdia	temperatura_c	umidade_relativa_percentual	focos_queimada
0	2016-01-01 00:00	183.4	9.4	16.9	0.9	0.0	27.6	85	NaN
1	2016-01-01 06:00	191.6	7.0	15.8	1.4	NaN	24.8	94	NaN
2	2016-01-01 12:00	192.8	7.0	20.8	0.6	NaN	27.8	86	NaN
3	2016-01-01 18:00	184.9	12.7	16.0	0.3	NaN	29.3	77	NaN
4	2016-01-02 00:00	201.0	7.0	16.8	1.1	3.0	25.7	92	NaN
...
8759	2015-12-30 18:00	246.9	27.2	21.8	0.2	NaN	29.5	70	NaN
8760	2015-12-31 00:00	289.5	14.4	25.2	1.4	0.0	26.2	88	NaN
8761	2015-12-31 06:00	297.1	20.7	28.3	0.9	NaN	24.8	88	NaN
8762	2015-12-31 12:00	226.0	14.6	20.9	0.6	NaN	28.0	77	NaN
8763	2015-12-31 18:00	171.7	16.2	17.3	0.2	NaN	28.9	78	NaN

8764 rows x 9 columns

fonte: Produção do próprio autor.

Figura 10 – Dados filtrados e tratados

```
df_air = df_air.groupby("datahora").mean().reset_index()
df_air
```

	datahora	co_ppb	o3_ppb	pm25_ugm3	so2_ugm3	precipitacao_mmdia	temperatura_c	umidade_relativa_percentual	focos_queimada
0	2013-01-01	135.400	11.125	8.175	0.875	0.50	26.450	81.75	0.0
1	2013-01-02	155.050	10.925	11.675	1.000	0.50	26.400	82.00	0.0
2	2013-01-03	137.400	12.225	8.800	0.825	0.00	27.400	81.25	0.0
3	2013-01-04	128.300	13.750	11.650	0.925	0.25	27.750	76.25	0.0
4	2013-01-05	128.250	10.950	7.800	1.075	2.25	26.625	84.75	0.0
...
2186	2018-12-27	78.375	3.750	7.850	0.750	0.00	26.000	96.50	0.0
2187	2018-12-28	66.450	4.625	6.100	0.350	7.00	26.000	96.25	0.0
2188	2018-12-29	66.275	4.675	6.050	0.300	0.50	25.700	96.25	0.0
2189	2018-12-30	59.925	5.700	4.750	0.300	1.00	25.575	95.50	0.0
2190	2018-12-31	63.225	3.375	6.775	0.475	0.25	25.050	96.50	0.0

2191 rows x 9 columns

fonte: Produção do próprio autor.

Após selecionar o botão enviar, uma lista de arquivos aparece para serem baixados e conforme a descrição acima foram retornados 72 arquivos. Os 6 primeiros aparecem na Figura 12.

Os arquivos estão em formato .dbc, esse formato impossibilita manipular os arquivos pela ferramenta Pandas, por isso, foi utilizado o *software* TABWIN disponibilizado pelo Datasus. Nesse programa os arquivos são primeiramente expandidos para .dbf, após isso é possível converter um arquivo por vez para .csv, possibilitando a manipulação em código. Observa-se os processos de conversão nas Figura 13 e Figura 14.

Agora, com os arquivos do Datasus em extensão csv realizou-se a filtragem das colunas de interesse, sendo as colunas, a data e o diagnóstico da internação filtrando para a cidade de Cuiabá. Fez-se a soma das internações por problemas respiratórios para cada dia. O resultado obtido pode ser observado na Figura 15.

Conforme apresentado na Figura 15 os diagnósticos de doenças possuem um código, para selecionar apenas as doenças respiratórias, verificou-se o índice de Classificação Internacional de Doenças (CID-10), no índice, observa-se que a seção "J" reúne os códigos das doenças respiratórias, dessa forma fez-se a seleção apenas dos códigos que começavam com a letra J na base do Datasus.

Para finalizar a obtenção dos dados foi realizada a junção das duas tabelas: dados do SISAM e diagnóstico das internações. A tabela resultante está apresentada na Figura 16.

Figura 11 – Interface de coleta de dados do Datasus

Download de arquivos

Fonte

SIASUS - Sistema de Informações Ambulatoriais do SUS
SIHSUS - Sistema de Informações Hospitalares do SUS
 SIM - Sistema de Informações de Mortalidade
 SINAN - Sistema de Informações de Agravos de Notificação

Modalidade

Arquivos auxiliares para tabulação
Dados
 Documentação

Tipo de Arquivo

ER - AIH Rejeitadas com código de erro
RD - AIH Reduzida
 RJ - AIH Rejeitadas
 SP - Serviços Profissionais

Ano

2016
2015
 2014
 2013
 2012

Mês

Setembro
 Outubro
 Novembro
 Dezembro

UF

MG
 MS
MT
 PA

Enviar

fonte: Datasus (2023)

Figura 12 – Arquivos para *download* do Datasus.

#		Fonte	Modalidade	Tipo de Arquivo
0	<input checked="" type="checkbox"/>	SIHSUS	Dados	RDMT1301.dbc
1	<input checked="" type="checkbox"/>	SIHSUS	Dados	RDMT1302.dbc
2	<input checked="" type="checkbox"/>	SIHSUS	Dados	RDMT1303.dbc
3	<input checked="" type="checkbox"/>	SIHSUS	Dados	RDMT1304.dbc
4	<input checked="" type="checkbox"/>	SIHSUS	Dados	RDMT1305.dbc
5	<input checked="" type="checkbox"/>	SIHSUS	Dados	RDMT1306.dbc

fonte: Datasus (2023)

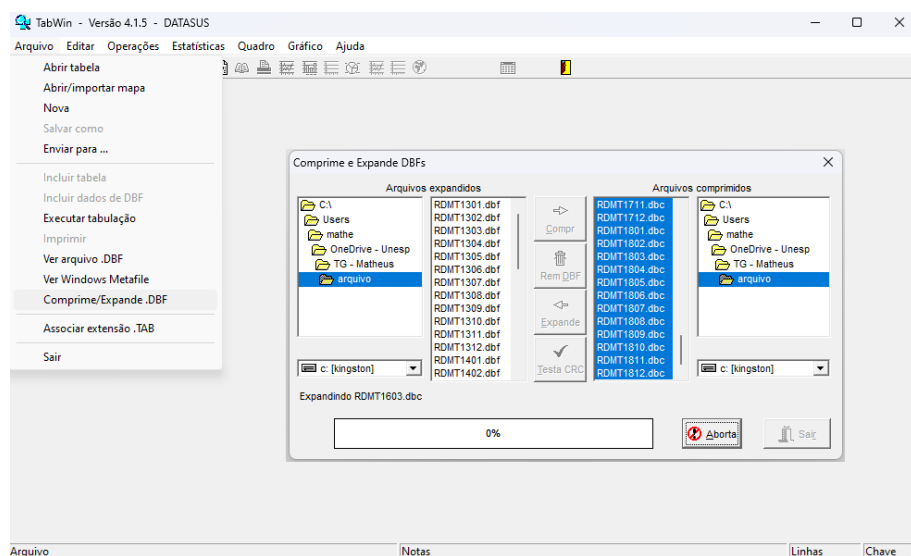
3.3 MODELAGEM DO PROBLEMA

A modelagem do problema foi dividida da em duas abordagens distintas, isso porque, como os dados são uma série temporal e o problema foi classificado como sendo de regressão, optou-se na primeira abordagem por utilizar os modelos de *machine learning* sem se importar com a característica temporal dos dados. A segunda abordagem, diferentemente da primeira, foi modelar uma rede LSTM tratando como importante a característica temporal dos dados para encontrar padrões na série e prever o número de pacientes internados por doenças respiratórias.

Em todos os modelos apenas 3 *features* foram utilizadas nos treinamentos, sendo elas: *o3_ppb*, *pm25_ugm3*, *temperatura_c*, *umidade_relativa_percentual*. A variação da quantidade de variáveis de entrada pode ser realizada em trabalhos futuros.

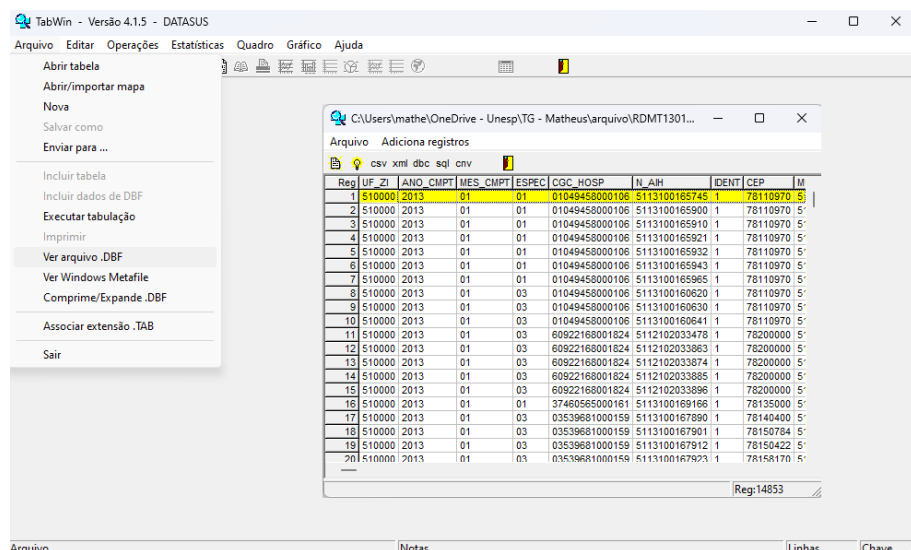
O objetivo principal é avaliar quais dos modelos desempenhará melhor e comparar as abordagens.

Figura 13 – Conversão dos arquivos em .dbf.



fonte: Produção do próprio autor.

Figura 14 – Conversão dos arquivos em .csv.



fonte: Produção do próprio autor.

3.3.1 Primeira abordagem

Antes de iniciar o treinamento dos modelos de *machine learning* é necessário fazer um tratamento dos dados a fim de maximizar o desempenho dos modelos.


O primeiro passo foi criar uma coluna apenas com o valor de cada ano, após isso, fez-se a divisão dos dados em 80% para treino e 20% para teste utilizando a função *"train_test_split"* da biblioteca *"sklearn"*, conforme apresentado na Figura 17 configurou-se o parâmetro *"stratify"* para estratificar os dados pela coluna com o valor dos anos. Isso para que o conjunto de treino e teste mesmo aleatorizados tenham a mesma quantidade de valores aleatórios por ano. Assim, evitar que um determinado ano por ter mais dados após realizar o *split* tenha mais influência no modelo do que os outros anos.

Uma análise de correlação linear entre as *features* e a variável de predição foi realizada a fim de identificar quais *features* tinha uma capacidade maior de informar uma tendencia na variável de saída.

Figura 15 – Tratamento final dos dados de doenças respiratórias.

DIAG_PRINC	data	respiratory
0	J189 2012-08-09	1
1	J152 2012-08-10	1
2	J159 2012-08-26	1
3	J449 2012-08-28	1
4	J985 2012-09-03	1
...
17347	J180 2018-12-13	1
17348	J449 2018-12-13	1
17349	J159 2018-12-18	1
17350	J189 2018-12-18	1
17351	J189 2018-12-23	1

17352 rows × 3 columns



date	respiratory
0 2013-01-01	6.0
1 2013-01-02	8.0
2 2013-01-03	12.0
3 2013-01-04	9.0
4 2013-01-05	10.0
...	...
2186 2018-12-27	0.0
2187 2018-12-28	0.0
2188 2018-12-29	0.0
2189 2018-12-30	0.0
2190 2018-12-31	0.0

2191 rows × 2 columns

fonte: Produção do próprio autor.

Figura 16 – Tabela com todas as *features* necessárias para os treinamentos dos modelos

date	co_ppb	o3_ppb	pm25_ugm3	so2_ugm3	temperatura_c	umidade_relativa_percentual	precipitacao_mmdia	focos_queimada	respiratory
0 2013-01-01	135.400	11.125	8.175	0.875	26.450	81.75	0.50	0.0	6.0
1 2013-01-02	155.050	10.925	11.675	1.000	26.400	82.00	0.50	0.0	8.0
2 2013-01-03	137.400	12.225	8.800	0.825	27.400	81.25	0.00	0.0	12.0
3 2013-01-04	128.300	13.750	11.650	0.925	27.750	76.25	0.25	0.0	9.0
4 2013-01-05	128.250	10.950	7.800	1.075	26.625	84.75	2.25	0.0	10.0
...
2186 2018-12-27	78.375	3.750	7.850	0.750	26.000	96.50	0.00	0.0	0.0
2187 2018-12-28	66.450	4.625	6.100	0.350	26.000	96.25	7.00	0.0	0.0
2188 2018-12-29	66.275	4.675	6.050	0.300	25.700	96.25	0.50	0.0	0.0
2189 2018-12-30	59.925	5.700	4.750	0.300	25.575	95.50	1.00	0.0	0.0
2190 2018-12-31	63.225	3.375	6.775	0.475	25.050	96.50	0.25	0.0	0.0

2191 rows × 10 columns

fonte: Produção do próprio autor.

Figura 17 – Comparação entre estratificar o *dataset* por ano e apenas randomizar

```

strat_train_set, strat_test_set = train_test_split(
    resp, test_size=0.2, stratify=resp["year"], random_state=42
)

```

✓ 0.6s

	Overall %	Stratified %	Random %	Strat. Error %	Rand. Error %
Year					
2013	16.66	16.63	17.77	-0.18	6.65
2014	16.66	16.63	15.72	-0.18	-5.65
2015	16.66	16.63	15.95	-0.18	-4.28
2016	16.70	16.86	14.81	0.91	-11.36
2017	16.66	16.63	18.91	-0.18	13.49
2018	16.66	16.63	16.86	-0.18	1.19

fonte: Produção do próprio autor.

Conforme apresentado na Tabela 4 nenhuma das variáveis de entrada apresentou correlação linear, isso não significa que elas não servem para prever o número de internações por doenças respiratórias, mas sim que a saída não varia linearmente com algumas das entradas.

A última etapa antes do treinamento é a construção de uma *pipeline* que tem a função de executar funções de transformações do *dataframe* de treino antes de ser aplicado no modelo. Foi feita a construção de uma *pipeline* simples com o "*StandardScaler*" que subtrai do valor de cada *feature*

Tabela 4 – Correlação linear entre as *features* e a variável de saída

<i>features</i>	correlação
so2_ugm3	0.029319
o3_ppb	-0.009726
temperatura_c	-0.028572
co_ppb	-0.047439
precipitacao_mmdia	-0.064309
pm25_ugm3	-0.091198
umidade_relativa_percentual	-0.171462

fonte: Produção do próprio autor.

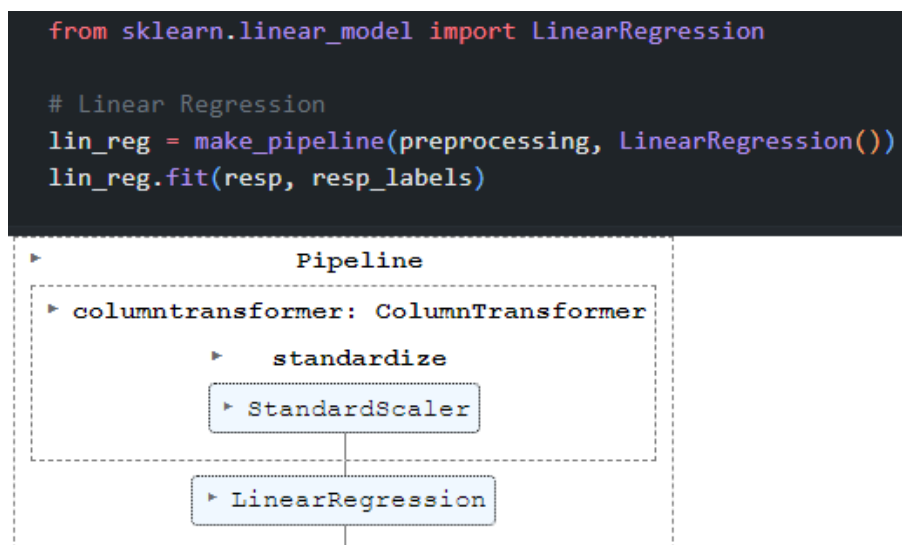
pela média e divide pelo desvio padrão, esse processo é conhecido como normalização e sendo um requerimento comum para os modelos de aprendizado de máquina.

3.3.1.1 Modelo de regressão linear

O modelo de regressão linear foi escolhido como base de comparação por ser um dos modelos mais simples e assim comparar com os estimadores mais sofisticados.

A API do *sklearn* é simples e intuitiva, a Figura 18 apresenta o código utilizado para a construção do modelo:

Figura 18 – Construção do modelo de regressão linear



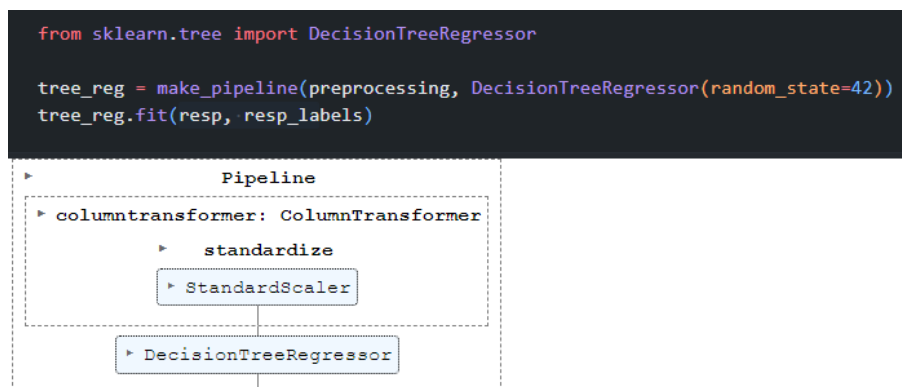
fonte: Produção do próprio autor.

3.3.1.2 Decision Tree

O *Decison Tree* foi utilizado por ser realmente mais robusto que a regressão linear e de acordo com (Géron, 2022) é capaz de encontrar complexas relações de correlações não lineares.

O código para a construção da árvore de decisão está na Figura 19:

Figura 19 – Construção do modelo de árvore de decisão



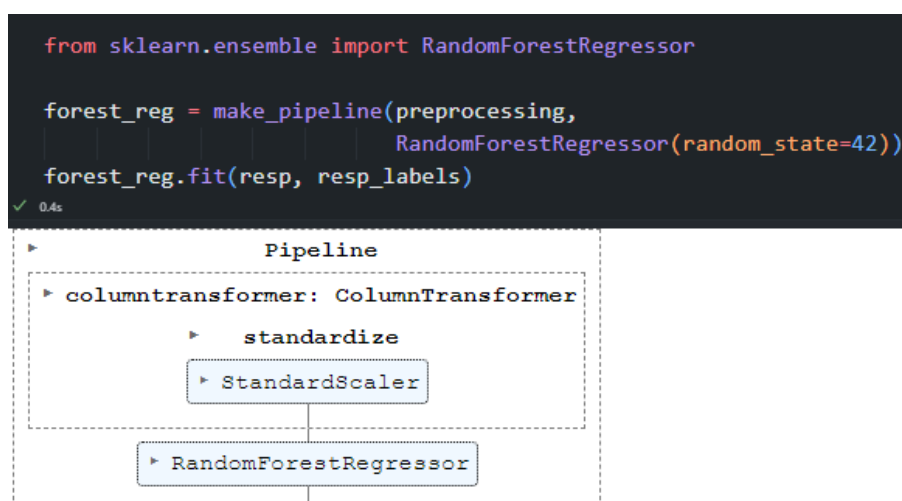
fonte: Produção do próprio autor.

3.3.1.3 Random Forest

O modelo de floresta aleatória como descrito no capítulo 2.2.2 tem um aumento de robustez e melhoria em alguns pontos relacionados a uma *Decision Tree* simples, o *Random Forest* faz um conjunto de unidades de árvores de decisão diminuindo a possibilidade de *overfitting*.

De maneira semelhante ao código anterior, a Figura 20 apresenta a construção do modelo:

Figura 20 – Construção do modelo de floresta aleatória



fonte: Produção do próprio autor.

3.3.2 Segunda abordagem

A segunda abordagem baseia-se na utilização dos dados arranjados como uma série temporal e para isso decidiu-se modelar uma RNN do tipo LSTM. As redes LSTM consomem dados em formato de série temporal, portanto nessa nova abordagem necessitou-se realizar novamente o tratamento da base de dados para essa aplicação.

Os dados foram ordenados por ordem cronológica e então fez-se a normalização utilizando a ferramenta "*MinMaxScaler*" do *framework* "*sklearn*". O normalizador utilizado foi configurado para transformar os dados entre 0 e 1 para que todos os dados tenham a mesma ordem de grandeza ajudando no desempenho da RNN.

Após a normalização realizou-se a transformação dos dados para se adequar a um problema de aprendizado supervisionado em séries históricas, dessa forma é possível considerar um *lag* da quantidade de dias desejados. Isso significa que condições climáticas e de qualidade do ar não provocam, necessariamente, uma complicação respiratória a uma pessoa no mesmo dia, mas pode levar a pessoa a se sentir mal e procurar ajuda médica dias depois. Considerando um *lag* de 3 dias, observa-se a transformação dos dados na Tabela 5.

3.3.2.1 LSTM

A construção do modelo da LSTM foi feita utilizando uma ferramenta de *tuning* dos hiperparâmetros, ou seja, uma ferramenta que busca os melhores hiperparâmetros automaticamente dentro dos limites estipulados por código em cada um dos hiperparâmetros.

Os dados foram divididos entre 80% para treino e 20% para teste, sendo que dos 80% do *dataset* treino, 10% foi utilizado para validação do modelo durante as épocas de treinamento.

O modelo foi construído conforme o código da Figura 21.

Figura 21 – Função de *tuning* para construção do modelo de LSTM

```
tf.compat.v1.keras.layers.CuDNNLSTM

def lstm_model_builder(hp):

    hp_learning_rate = hp.Choice('learning_rate', values=['1e-2', '1e-3', '1e-4'])

    model = Sequential()

    model.add(LSTM(hp.Int('input_unit',
                           min_value=4,
                           max_value=512,
                           step=8),
                    return_sequences=False,
                    input_shape=(train_x.shape[1],train_x.shape[2])
                    ))

    model.add(Dense(train_y.shape[1], activation=hp.Choice('dense_activation',values=['relu', 'sigmoid'],default='relu'))))
    model.compile(loss='mean_squared_error', optimizer='adam',metrics = ['mse', 'mae'])

    return model
```

fonte: Produção do próprio autor.

A ferramenta irá variar a taxa de aprendizagem do modelo, ou seja, quão rápido o modelo convergirá para o mínimo da função de perda. Com relação a estrutura da LSTM, será variado a quantidade de unidades LSTM entre 4 e 512 em múltiplos de 8 aleatoriamente e por fim, a função de ativação da saída do modelo poderá ser *relu* (*rectified linear unit*) ou sigmoidal.

Tabela 5 – Transformação dos dados para aprendizado supervisionado (lag=3 dias)

var1(t-3)	var2(t-3)	var3(t-3)	var1(t-2)	var2(t-2)	var3(t-2)	var1(t-1)	var2(t-1)	var3(t-1)	var1(t)	var2(t)	var3(t)
0.025462	0.746212	0.214286	0.037227	0.750000	0.285714	0.027563	0.738636	0.428571	0.037143	0.662879	0.321429
0.037227	0.750000	0.285714	0.027563	0.738636	0.428571	0.037143	0.662879	0.321429	0.024202	0.791667	0.357143
0.027563	0.738636	0.428571	0.037143	0.662879	0.321429	0.024202	0.791667	0.357143	0.017563	0.772727	0.142857
0.037143	0.662879	0.321429	0.024202	0.791667	0.357143	0.017563	0.772727	0.142857	0.018655	0.791667	0.214286
0.024202	0.791667	0.357143	0.017563	0.772727	0.142857	0.018655	0.791667	0.214286	0.025210	0.803030	0.250000
0.017563	0.772727	0.142857	0.018655	0.791667	0.214286	0.025210	0.803030	0.250000	0.036134	0.833333	0.500000
0.018655	0.791667	0.214286	0.025210	0.803030	0.250000	0.036134	0.833333	0.500000	0.035630	0.837121	0.142857
0.025210	0.803030	0.250000	0.036134	0.833333	0.500000	0.035630	0.837121	0.142857	0.032269	0.750000	0.142857
0.036134	0.833333	0.500000	0.035630	0.837121	0.142857	0.032269	0.750000	0.142857	0.028403	0.821970	0.214286
0.035630	0.837121	0.142857	0.032269	0.750000	0.142857	0.028403	0.821970	0.214286	0.020000	0.871212	0.285714
0.032269	0.750000	0.142857	0.028403	0.821970	0.214286	0.020000	0.871212	0.285714	0.007731	0.859848	0.178571
0.028403	0.821970	0.214286	0.020000	0.871212	0.285714	0.007731	0.859848	0.178571	0.014706	0.814394	0.250000
0.020000	0.871212	0.285714	0.007731	0.859848	0.178571	0.014706	0.814394	0.250000	0.023109	0.704545	0.214286
0.007731	0.859848	0.178571	0.014706	0.814394	0.250000	0.023109	0.704545	0.214286	0.023529	0.776515	0.250000
0.014706	0.814394	0.250000	0.023109	0.704545	0.214286	0.023529	0.776515	0.250000	0.031008	0.784091	0.250000
0.023109	0.704545	0.214286	0.023529	0.776515	0.250000	0.031008	0.784091	0.250000	0.027899	0.753788	0.107143
0.023529	0.776515	0.250000	0.031008	0.784091	0.250000	0.027899	0.753788	0.107143	0.052773	0.886364	0.071429
0.031008	0.784091	0.250000	0.027899	0.753788	0.107143	0.052773	0.886364	0.071429	0.039832	0.844697	0.285714
0.027899	0.753788	0.107143	0.052773	0.886364	0.071429	0.039832	0.844697	0.285714	0.026555	0.833333	0.285714
0.052773	0.886364	0.071429	0.039832	0.844697	0.285714	0.026555	0.833333	0.285714	0.019076	0.829545	0.250000

fonte: Produção do próprio autor.

4 ANÁLISE DOS RESULTADOS OBTIDOS

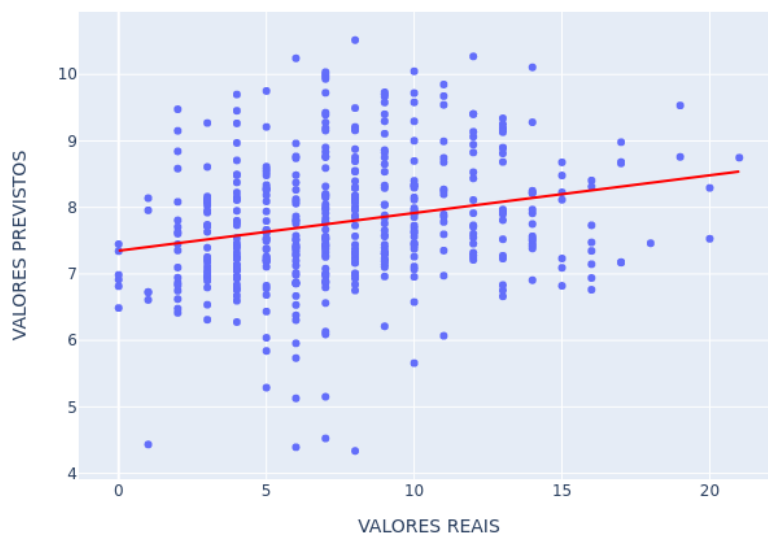
Neste capítulo será apresentado os resultados obtidos por cada um dos modelos, realizando uma comparação e apontando o modelo que obteve os melhores resultados.

- Modelo de Regressão Linear

Conforme apresentado na Tabela 6 o modelo obteve um RMSE de aproximadamente 15 internações por doenças respiratórias, isso significa que, para cada internado o modelo estipulou em média 15 internados a mais ou a menos. Esse resultado era esperado, pois foi verificado na Tabela 4 que não existe correlação linear com a variável de saída.

A Figura 22 apresenta a relação entre os dados de teste dos valores reais de internações e os valores previstos.

Figura 22 – Diagrama de dispersão dos dados reais e dos previstos do modelo LR



fonte: Produção do próprio autor.

- Árvore de Decisão

Esse modelo obteve uma grande melhora de performance quando comparado ao modelo de regressão linear, uma das diferenças foi a realização da validação cruzada. Observando a tabela 7 nota-se que o RMSE diminuiu 2,74 vezes.

Tabela 6 – Métricas do modelo de regressão linear

rmse	mae	mape
14.922755	3.048696	4.307988e+14

fonte: Produção do próprio autor.

Tabela 7 – Métricas do modelo de árvore de decisão

rmse	mae	mape
5.344163	4.155221	2.620422e+14

fonte: Produção do próprio autor.

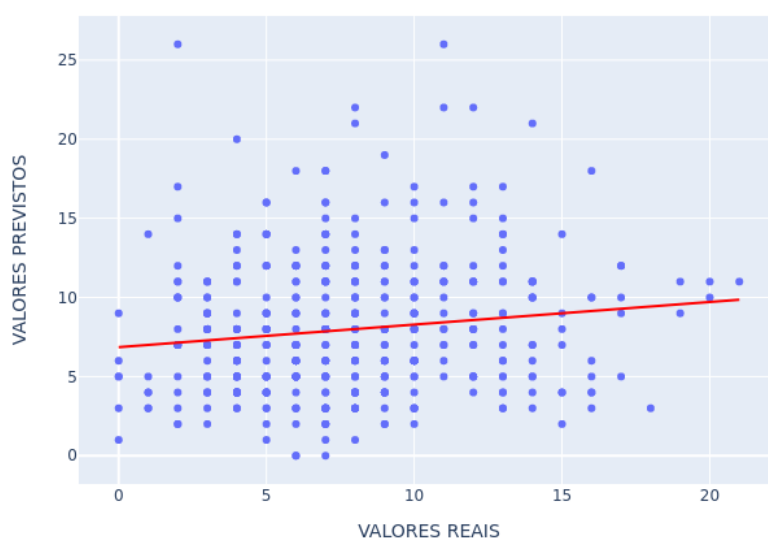
Tabela 8 – Métricas do modelo de floresta aleatória

rmse	mae	mape
3.952855	3.088132	2.852175e+14

fonte: Produção do próprio autor.

A Figura 23 apresenta a relação entre os dados de teste dos valores reais de internações e os valores previstos.

Figura 23 – Diagrama de dispersão dos dados reais e dos previstos do modelo DT



fonte: Produção do próprio autor.

- Floresta Aleatória

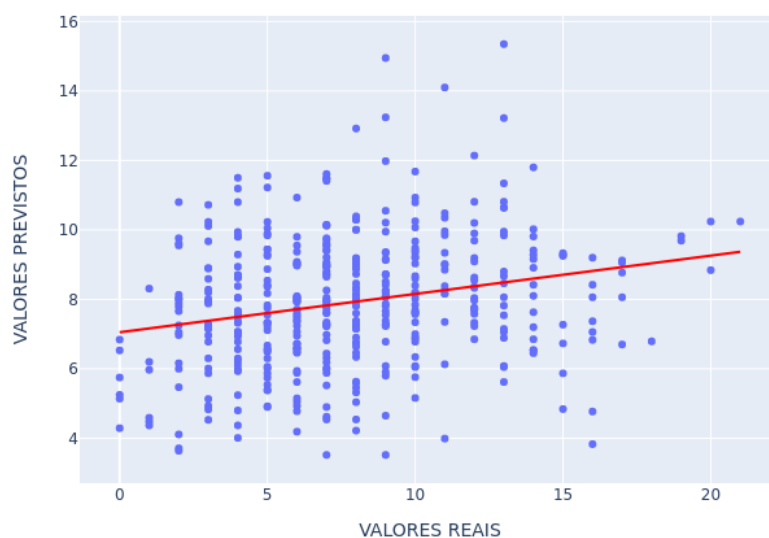
Esse modelo de *machine learning* é o mais robusto dos testados da primeira abordagem, pela tabela 8 pode-se concluir que o RF obteve o melhor resultado entre o LR e DT, o RMSE encontrado foi de aproximadamente 4 internados a mais ou a menos para cada valor real.

A Figura 24 apresenta a relação entre os dados de teste dos valores reais de internações e os valores previstos.

- LSTM

O modelo de LSTM representa uma grande diferença no tratamento de resolução do problema, pois o trata como uma regressão de séries temporais. Isso pode permitir encontrar padrões temporais na série que não foram observados nos modelos de *machine learning* descritos anteriormente.

Figura 24 – Diagrama de dispersão dos dados reais e dos previstos do modelo RF



fonte: Produção do próprio autor.

Tabela 9 – Métricas do modelo LSTM

rmse	mae	mape
3.239208	2.720342	1.503353e+15

fonte: Produção do próprio autor.

Analizando a Tabela 9 percebe-se que o modelo obteve o menor RMSE entre todos os modelos testados, isso mostra a robustez de uma rede neural recorrente em encontrar padrões não lineares.

A Figura 25 apresenta a relação entre os dados de teste dos valores reais de interações e os valores previstos.

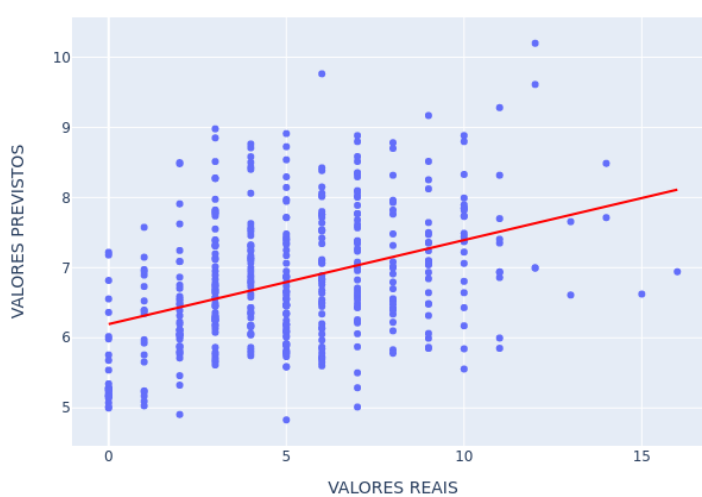
A Tabela 10 apresenta os resultados obtidos para cada um dos modelos:

Tabela 10 – Resultados dos modelos

Modelo	rmse	MAE	MAPE
Regressão Linear	14.922755	3.048696	4.307988e+14
Árvore de Decisão	5.344163	4.155221	2.620422e+14
Árvore Aleatória	3.952855	3.088132	2.852175e+14
LSTM	3.239208	2.720342	1.503353e+15

fonte: Produção do próprio autor.

Figura 25 – Diagrama de dispersão dos dados reais e dos previstos do modelo LSTM



fonte: Produção do próprio autor.

5 CONCLUSÃO

O principal objetivo desse trabalho foi comparar as técnicas de *machine learning* e rede neural artificial apresentadas e verificar qual modelo obteve as previsões mais próximas dos valores reais observados.

Analisando os resultados dos modelos pela métrica MAPE, observa-se que os valores obtidos são extremamente altos. Isso ocorreu, pois muitos valores de internações observadas são iguais a 0 e dessa forma, ao calcular esse erro ocorre uma divisão por 0, resultando em valores enormes. Por isso, o MAPE não pode ser levado em consideração na análise.

Com relação ao erro médio absoluto (MAE) o modelo que obteve a melhor performance foi a LSTM, no entanto, o RF e LR apresentaram erros muito próximos, com leve vantagem para a LR. Analisando pela raiz do erro quadrático médio (RMSE), observa-se que o *Random Forest* desempenhou desse vez muito melhor que o modelo de regressão linear, isso significa que a LR possui erros relativamente maiores que o RF. Novamente, a LSTM foi o modelo que apresentou o menor erro.

Dessa forma, tomando como principal métrica de avaliação a raiz do erro quadrático médio (RMSE). O melhor modelo foi um tipo de rede neural recorrente, a *Long Short Term Memory* (LSTM). A LSTM foi mais eficaz em encontrar correlações não lineares nos dados de entrada com a variável de saída.

Assim, diante da problemática apresentada nos capítulos iniciais, conclui-se que esse trabalho contribui para a geração de novas informações adequadas que podem ajudar líderes e gestores dos sistemas de saúde da cidade de Cuiabá-MT e de todo o território nacional a traçar estratégias de melhoria dos serviços de saúde e gestão dos recursos, principalmente, mediante um contexto de variações na demanda nos sistemas de saúde. Promovendo, melhor qualidade de vida e condições ao combate das doenças respiratórias.

5.1 PROPOSTAS PARA FUTURAS PESQUISAS

- Aumentar a base de dados;
- Analisar a influência das *features* não utilizadas, variando-as na base de dados;
- Analisar a influência da utilização de diferentes *lags*, variando entre 0 a 7;
- Realizar *Hyperparameter boost* da *Random Forest*;
- Utilizar outros modelos de *machine learning* como XGboost e SVR;
- Aumentar a complexidade do modelo de LSTM.

REFERÊNCIAS

- AGENCY, I. E. **Energy Statistics Data Browser**. 2023. Disponível em: <<https://www.iea.org/data-and-statistics/data-tools/energy-statistics-data-browser?country=WORLD&fuel=Energy%20supply&indicator=TESbySource>>. Acesso em: 21 de jan. de 2023.
- CARBONNELLE, P. **PYPL PopularitY of Programming Language**. 2023. Disponível em: <<https://pypl.github.io/PYPL.html>>. Acesso em: 21 de jan. de 2023.
- GROSSO, G. do M. **Queimadas deixam situação do ar crítica em Cuiabá**. 2017. Disponível em: <<http://www.mt.gov.br/-/8185483-queimadas-deixam-situacao-do-ar-critica-em-cuiaba>>. Acesso em: 24 de jan. de 2023.
- GÉRON, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd Edition**. Sebastopol, Califórnia - EUA: O'REILLY, 2022.
- IBGE. **Cidades e Estados: Cuiabá**. 2023. Disponível em: <<https://www.ibge.gov.br/cidades-e-estados/mt/cuiaba.html>>. Acesso em: 23 de jan. de 2023.
- IBGE. **Cuiabá**. 2023. Disponível em: <<https://cidades.ibge.gov.br/brasil/mt/cuiaba/panorama>>. Acesso em: 23 de jan. de 2023.
- LATORRE, M. do Rosário Dias de O.; CARDOSO, M. R. A. Análise de séries temporais em epidemiologia: uma introdução sobre os aspectos metodológicos. **Rev. Bras. Epidemiol.**, v. 4, n. 147, p. 145 – 152, 2001.
- LUTZ, M. **Learning Python, 5th Edition**. Sebastopol, Califórnia - EUA: O'REILLY, 2013.
- PANCH, T.; SZOLOVITS, P.; ATUN, R. Artificial intelligence, machine learning and health systems. **Journal of Global Health**, v. 8, n. 8, p. 1 – 8, 2018.
- SOYIRI, I. N.; REIDPATH, D. D.; SARRAN, C. Forecasting peak asthma admissions in london: an application of quantile regression models. **International Journal of Biometeorology**, v. 1, n. 2, p. 1 – 11, 2012.
- STARMER, S. with J. **Long Short-Term Memory (LSTM), claramente explicado**. 2022. Disponível em: <https://www.youtube.com/watch?v=YCzL96nL7j0&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=84&t=580s>. Acesso em: 25 de jan. de 2023.
- UNIDAS, O. das N. **Novas diretrizes da OMS sobre qualidade do ar reduzem valores seguros para poluição**. 2021. Disponível em: <<https://brasil.un.org/pt-br/145721-novas-diretrizes-da-oms-sobre-qualidade-do-ar-reduzem-valores-seguros-para-poluicao>>. Acesso em: 18 de jan. de 2023.
- UNIDAS, O. das N. **Temperatura média global tem 50exceder 1,5°C até 2026**. 2022. Disponível em: <<https://brasil.un.org/pt-br/181236-temperatura-media-global-tem-50-de-chance-de-exceder-15degc-ate-2026#:~:text=De%20acordo%20com%20os%20novos,%2C%20saltou%20para%20quase%2050%25.>>> Acesso em: 20 de jan. de 2023.
- WHO. **Health service delivery**. Genève, Switzerland: World Health Organization, 2010.
- XIAO, P. **Artificial Intelligence Programming with Python: From Zero to Hero**. Hoboken, Nova Jersey - EUA: Wiley, 2022.

XU, Z.; HU, W.; TONG, S. Temperature variability and childhood pneumonia: An ecological study. **Environmental Health**, v. 1, n. 8, p. 1 – 8, 2014.