RESSALVA

Atendendo solicitação do(a) autor(a), o texto completo desta tese será disponibilizado somente a partir de 28/01/2027.

UNESP – UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
CÂMPUS DE PRESIDENTE PRUDENTE
Programa de Pós-graduação em Ciências Cartográficas

RENATA CRISTINA FAUSTINO

ANÁLISE ESPAÇO-TEMPORAL E MODELAGEM PREDITIVA DOS FATORES ASSOCIADOS À DISSEMINAÇÃO DE BACTÉRIAS MULTIRRESISTENTES NO ESTADO DE SÃO PAULO

Presidente Prudente – SP
2025

UNESP – UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
CÂMPUS DE PRESIDENTE PRUDENTE
Programa de Pós-graduação em Ciências Cartográficas

RENATA CRISTINA FAUSTINO

ANÁLISE ESPAÇO-TEMPORAL E MODELAGEM PREDITIVA DOS FATORES ASSOCIADOS À DISSEMINAÇÃO DE BACTÉRIAS MULTIRRESISTENTES NO ESTADO DE SÃO PAULO

Relatório de defesa de Doutorado apresentado ao Programa de Pós-Graduação em Ciências Cartográficas (PPGCC), da Faculdade de Ciências e Tecnologia (FCT), Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), campus de Presidente Prudente/SP, como parte dos requisitos para a obtenção do título de Doutora em Ciências Cartográficas.

Orientador: Prof. Dr. Edmur Azevedo Pugliesi
Coorientador: Prof. Dr. Carlos Magno Castelo Branco Fortaleza

Presidente Prudente – SP
2025

F268a

Faustino, Renata Cristina
Análise espaço-temporal e modelagem preditiva dos fatores associados à disseminação de bactérias multirresistentes no estado de São Paulo / Renata Cristina Faustino. -- Presidente Prudente, 2025
243 p. : il., tabs., mapas

Tese (doutorado) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências e Tecnologia, Presidente Prudente
Orientador: Edmur Azevedo Pugliesi
Coorientador: Carlos Magno Castelo Branco Fortaleza

1. Análise espacial. 2. Machine Learning. 3. Resistência Bacteriana. 4. Infecção hospitalar. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

IMPACTO POTENCIAL DESTA PESQUISA

Esta pesquisa investiga a proliferação de bactérias multirresistentes e seus fatores associados, utilizando duas abordagens distintas no estado de São Paulo, entre 2011 e 2019: a análise da distribuição espacial que identifica os locais mais afetados, e a modelagem preditiva, que estima a contribuição relativa de cada fator associado. Os resultados têm potencial para subsidiar políticas públicas voltadas ao aprimoramento do controle e monitoramento da disseminação de patógenos resistentes em hospitais, principalmente aqueles que dispõem de leitos de UTI.

POTENCIAL IMPACT OF THIS RESEARCH

This research investigates the proliferation of multidrug-resistant bacteria and their associated factors using two distinct approaches in the state of São Paulo, between 2011 and 2019: spatial distribution analysis, which identifies the most affected areas, and predictive modeling, which estimates the relative contribution of each associated factor. The results have the potential to support public policies aimed at improving the control and monitoring of the spread of resistant pathogens in hospitals, especially those with intensive care unit (ICU) beds.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Presidente Prudente

ATA DA DEFESA PÚBLICA DA TESE DE DOUTORADO DE RENATA CRISTINA FAUSTINO, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS CARTOGRÁFICAS, DA FACULDADE DE CIÊNCIAS E TECNOLOGIA - CÂMPUS DE PRESIDENTE PRUDENTE.

Aos 28 dias do mês de julho do ano de 2025, às 8h, por meio de Videoconferência, realizou-se a defesa de TESE DE DOUTORADO de RENATA CRISTINA FAUSTINO, intitulada **ANÁLISE ESPAÇO-TEMPORAL E MODELAGEM PREDITIVA DOS FATORES ASSOCIADOS À DISSEMINAÇÃO DE BACTÉRIAS MULTIRRESISTENTES NO ESTADO DE SÃO PAULO**. A Comissão Examinadora foi constituída pelos seguintes membros: Prof. Dr. EDMUR AZEVEDO PUGLIESI (Orientador(a) - Participação Presencial) do(a) Departamento de Cartografia / Faculdade de Ciencias e Tecnologia de Presidente Prudente, Prof. Dr. ROGÉRIO EDUARDO GARCIA (Participação Presencial) do(a) Universidade Estadual Paulista "Júlio de Mesquita Filho" - Faculdade de Ciências e Tecnologia, Prof. Dr. THOMAS NOGUEIRA VILCHES (Participação Virtual) do(a) Departamento de Biodiversidade e Bioestatística / Universidade Estadual Paulista "Júlio de Mesquita Filho", Prof. Dr. ÍCARO BOSZCZOWSKI (Participação Virtual) do(a) Subcomissão de Controle de Infecção Hospitalar / Instituto Central do Hospital das Clínicas - FMUSP, Prof. Dr. LUIZ EURIBEL PRESTES CARNEIRO (Participação Presencial) do(a) Universidade do Oeste Paulista - UNOESTE . Após a exposição pela doutoranda e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, a discente recebeu o conceito final: APROVADO . Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.

Prof. Dr. EDMUR AZEVEDO PUGLIESI

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Presidente Prudente

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: **ANÁLISE ESPAÇO-TEMPORAL E MODELAGEM PREDITIVA DOS FATORES ASSOCIADOS À DISSEMINAÇÃO DE BACTÉRIAS MULTIRRESISTENTES NO ESTADO DE SÃO PAULO**

AUTORA: RENATA CRISTINA FAUSTINO
ORIENTADOR: EDMUR AZEVEDO PUGLIESI
COORIENTADOR: CARLOS MAGNO CASTELO BRANCO FORTALEZA

Aprovada como parte das exigências para obtenção do Título de Doutora em Ciências Cartográficas, área: Aquisição, Análise e Representação de Informações Espaciais pela Comissão Examinadora:

gov.br Documento assinado digitalmente
EDMUR AZEVEDO PUGLIESI
Data: 05/08/2025 20:31:51-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. EDMUR AZEVEDO PUGLIESI (Participaçao Presencial)
Departamento de Cartografia / Faculdade de Ciencias e Tecnologia de Presidente Prudente

gov.br Documento assinado digitalmente
ROGERIO EDUARDO GARCIA
Data: 07/08/2025 11:39:00-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. ROGÉRIO EDUARDO GARCIA (Participaçao Presencial)
Universidade Estadual Paulista "Júlio de Mesquita Filho" - Faculdade de Ciências e Tecnologia

gov.br Documento assinado digitalmente
THOMAS NOGUEIRA VILCHES
Data: 06/08/2025 08:12:04-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. THOMAS NOGUEIRA VILCHES (Participaçao Virtual)
Departamento de Biodiversidade e Bioestatística / Universidade Estadual Paulista "Júlio de Mesquita Filho"

gov.br Documento assinado digitalmente
ICARO BOSZCZOWSKI
Data: 10/08/2025 18:14:47-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. ÍCARO BOSZCZOWSKI (Participaçao Virtual)
Subcomissão de Controle de Infecção Hospitalar / Instituto Central do Hospital das Clínicas - FMUSP

gov.br Documento assinado digitalmente
LUIZ EURIBEL PRESTES CARNEIRO
Data: 17/08/2025 18:49:51-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. LUIZ EURIBEL PRESTES CARNEIRO (Participaçao Presencial)
Universidade do Oeste Paulista - UNOESTE

Presidente Prudente, 28 de julho de 2025

À minha querida mãezinha Nilsa (in memoriam), cujo amor incondicional, força, coragem, sabedoria e alegria continuam a iluminar meu caminho e a inspirar cada passo da minha jornada. Para sempre em meu coração.

AGRADECIMENTOS

Primeiramente, agradeço a Deus, que me deu forças diárias para enfrentar essa longa jornada chamada doutorado.

À minha família e aos amigos que, mesmo já tendo decorado a frase “não posso, tenho que escrever a tese de doutorado”, nunca deixaram de me apoiar, deixo aqui meu sincero agradecimento. Agradeço especialmente ao meu pai, Carlos Faustino, à minha tia, Elisabete Dias, e ao meu namorado, Renan Escobar, que acompanharam de perto cada capítulo, cada madrugada em claro.

Sou igualmente grata aos amigos que, com frequência, me perguntam quando será a defesa, me convidam para sair e viajar, e mesmo diante das minhas negativas constantes, continuam firmes, escutando meus desabafos com carinho e paciência. Destaco com carinho Nádia Haga, Nathalia Rosa, Celi Muniz, Thaís Helena, Juliana Ribeiro, Igor Redivo, Mirian Miotto, Meire Miotto, Bruna Ribeiro, Eliana Francisquete e o pessoal do Sesc Thermas, pelos momentos de leveza e descontração, meu muito obrigada a todos vocês, sobretudo aos professores Emerson Silva e Daniela Reis. Agradeço também à Camerata, em especial ao maestro Luiz Antônio Perez Filho, por sempre acreditar em mim e repetir, com convicção, que sou capaz.

Ao meu orientador, Prof. Edmur Azevedo Pugliesi, por embarcar comigo nesse desafio. Ao meu coorientador, Prof. Carlos Magno Fortaleza, pela contribuição na definição do tema e abordagem do estudo. E ao Prof. Rogério Garcia, meu muito obrigada pelas ideias, códigos e socorros técnicos.

Aos colegas da pós-graduação e aos funcionários da UNESP, obrigada pelas conversas nos corredores, pelas trocas de ideias e pelo apoio que, muitas vezes, veio em forma de um café, um meme ou uma simples escuta.

A todos vocês, minha sincera gratidão por fazerem parte dessa caminhada.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

“Sê forte e corajoso; não temas, nem te espantes; porque o Senhor teu Deus é contigo, por onde quer que andares.”

(Josué 1:9, Bíblia Sagrada)

RESUMO

As infecções relacionadas à assistência à saúde são aquelas adquiridas durante a permanência do paciente em ambiente hospitalar. Nesses locais, pacientes infectados ou colonizados podem contaminar superfícies e equipamentos, facilitando a transmissão bacteriana entre pacientes, profissionais e unidades de saúde. Nesse contexto, o surgimento e a disseminação de resistência ao antibiótico carbapenêmico representam uma crescente ameaça à saúde pública global. Surtos esporádicos ou situações endêmicas envolvendo *Acinetobacter spp.* e enterobactérias resistentes aos carbapenêmicos têm sido cada vez mais reportados, com registros no Brasil a partir de 2011. Esta tese teve como objetivo principal analisar a disseminação espaço-temporal das infecções por bactérias multirresistentes associadas a fatores socioeconômicos e hospitalares, especificamente *Acinetobacter spp.* e enterobactérias, no estado de São Paulo, no período entre 2011 e 2019. Para isso, foram aplicadas técnicas de análise espacial e modelagem preditiva com algoritmos de *machine learning*, destacando a relevância desse estudo frente à ameaça global crescente da resistência antimicrobiana, enfatizada por organismos como a Organização Mundial da Saúde (OMS) e a Agência Nacional de Vigilância Sanitária (ANVISA). Na primeira etapa, de análise espacial, a metodologia envolveu análise descritiva de dados, técnicas de autocorrelação espacial univariada e bivariada e a elipse do desvio padrão. Os resultados obtidos revelaram tendências distintas para os dois grupos bacterianos estudados: enquanto as infecções por *Acinetobacter spp.* mostraram tendência geral de redução ao longo dos anos analisados, as infecções por enterobactérias aumentaram até 2016, apresentando posteriormente variações. Municípios pertencentes aos DRS da Grande São Paulo, Campinas e Baixada Santista destacaram-se pela presença de aglomerados espaciais positivos com valores elevados, sugerindo que hospitais com maior capacidade de internação, especialmente em unidades de terapia intensiva (UTIs), concentram taxas mais elevadas de infecções por *Acinetobacter spp.* e enterobactérias. Observa-se, ainda, que regiões com maior fluxo de pacientes estão associadas a uma disseminação mais intensa dessas infecções, sendo a variável paciente-dia um fator relevante, indicando que a alta rotatividade e o fluxo intenso de pacientes potencializam a disseminação bacteriana. Na segunda, voltada à modelagem preditiva, foram empregados os algoritmos de *machine learning*: Floresta Aleatória, Redes Neurais, XGBoost, LightGBM e CatBoost, em tarefas de regressão e classificação, com implementação em linguagem Python. As análises foram conduzidas separadamente para as unidades hospitalares e para os municípios. No nível municipal, a Floresta Aleatória apresentou o melhor desempenho nos modelos de regressão, com coeficiente de determinação (R^2) de 0,953 para predição de infecções por *Acinetobacter spp.*, enquanto o algoritmo Catboost apresentou o melhor desempenho para infecções por enterobatérias com R^2 de 0,984. As variáveis mais relevantes no modelo de Floresta Aleatória foram o número de leitos hospitalares, o total de pacientes-dia e o número de leitos de UTI. Para predição de infecções por enterobactérias, o CatBoost identificou o número de hospitais como variável mais relevante, seguido por leitos de UTI e leitos hospitalares. Ao considerar as unidades hospitalares, os modelos de regressão apresentaram resultados insatisfatórios obtendo um R^2 abaixo de 0,452. Já nos modelos de classificação, o algoritmo CatBoost destacou-se no nível municipal com área sob a curva ROC (AUC-ROC) de 0,953 para a predição de infecções por *Acinetobacter spp.* e de 0,984 para enterobactérias, enquanto os demais algoritmos também apresentaram bom desempenho, todos com AUC-ROC superior a 0,884. As

variáveis explicativas mais relevantes nesse contexto foram o número de leitos hospitalares, leitos de UTI, pacientes-dia e quantidade de hospitais. No nível hospitalar, o XGBoost apresentou o melhor desempenho na tarefa de classificação, com AUC-ROC de 0,866 para *Acinetobacter spp.* e de 0,812 para enterobactérias, sendo o número de leitos hospitalares e o total de pacientes-dia as variáveis de maior importância. Os resultados evidenciam que a disseminação das bactérias resistentes está fortemente relacionada à infraestrutura hospitalar, reforçando que indicadores como paciente-dia, número de leitos hospitalares e leitos de UTI são fundamentais para compreender e predizer a dinâmica epidemiológica das bactérias multirresistentes.

Palavras-chave: análise espacial; infecção hospitalar; *Acinetobacter spp.*; enterobactérias; *machine learning*.

ABSTRACT

Healthcare-associated infections are those acquired during a patient's stay in a hospital environment. In such settings, infected or colonized patients may contaminate surfaces and equipment, facilitating bacterial transmission among patients, healthcare workers, and healthcare facilities. In this context, the emergence and spread of resistance to carbapenem antibiotics represent an increasing threat to global public health. Sporadic outbreaks or endemic situations involving Acinetobacter spp. and carbapenem-resistant Enterobacteriaceae have been increasingly reported, with records in Brazil dating back to 2011. The main objective of this thesis was to analyze the spatiotemporal spread of multidrug-resistant bacterial infections associated with socioeconomic and hospital-related factors, specifically Acinetobacter spp. and Enterobacteriaceae, in the state of São Paulo from 2011 to 2019. To this end, spatial analysis techniques and predictive modeling using machine learning algorithms were applied, highlighting the relevance of this study in light of the growing global threat of antimicrobial resistance, as emphasized by organizations such as the World Health Organization (WHO) and the Brazilian Health Regulatory Agency (ANVISA). In the first stage, dedicated to spatial analysis, the methodology included descriptive data analysis, univariate and bivariate spatial autocorrelation techniques, and standard deviation ellipse. The results revealed distinct trends for the two bacterial groups studied: while infections caused by Acinetobacter spp. showed a general downward trend over the analyzed years, Enterobacteriaceae infections increased until 2016, followed by fluctuations. Municipalities belonging to the Regional Health Departments (DRS) of Greater São Paulo, Campinas, and Baixada Santista stood out for the presence of high-value positive spatial clusters, suggesting that hospitals with greater admission capacity, especially in intensive care units (ICUs), concentrate higher infection rates of both Acinetobacter spp. and Enterobacteriaceae. Furthermore, regions with higher patient flow were associated with more intense dissemination of these infections, with the variable “patient-day” emerging as a relevant factor, indicating that high turnover and intensive patient flow potentiate bacterial spread. In the second stage, focused on predictive modeling, the following machine learning algorithms were employed: Random Forest, Neural Networks, XGBoost, LightGBM, and CatBoost, in regression and classification tasks, implemented in Python. Analyses were conducted separately for hospital units and municipalities. At the municipal level, Random Forest showed the best performance in regression models, with a coefficient of determination (R^2) of 0.953 for predicting Acinetobacter spp. infections, while the CatBoost algorithm performed best for Enterobacteriaceae, with an R^2 of 0.984. The most relevant variables in the Random Forest model were the number of hospital beds, total patient-days, and number of ICU beds. For predicting Enterobacteriaceae infections, CatBoost identified the number of hospitals as the most relevant variable, followed by ICU beds and hospital beds. In hospital-level analyses, regression models yielded unsatisfactory results, with R^2 values below 0.452. However, in classification models, CatBoost stood out at the municipal level, with an area under the ROC curve (AUC-ROC) of 0.953 for Acinetobacter spp. and 0.984 for Enterobacteriaceae, while the other algorithms also performed well, all achieving AUC-ROC values above 0.884. The most relevant explanatory variables in this context were the number of hospital beds, ICU beds, patient-days, and number of hospitals. At the hospital level, XGBoost showed the best classification performance, with an AUC-ROC of 0.866 for Acinetobacter spp. and 0.812 for Enterobacteriaceae, with the number of hospital beds and total patient-days being the most important variables. The results clearly

demonstrate that the spread of resistant bacteria is strongly related to hospital infrastructure, reinforcing that indicators such as patient-days, number of hospital beds, and ICU capacity are essential to understand and predict the epidemiological dynamics of multidrug-resistant bacteria.

Keywords: spatial analysis; hospital infection; *Acinetobacter spp.*; *Enterobacteriaceae*; machine learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

AUC – A*rea Under the ROC curve*

CatBoost – *Categorical Boosting*

CNES – Cadastro Nacional de Estabelecimentos de Saúde

CRAB – *Acinetobacter* baumannii Resistentes a *Carbapenem*

CRE – Enterobactérias Resistente a *Carbapenem*

CVE – Centro de Vigilância Epidemiológica do Estado de São Paulo

DRS – Departamentos Regionais de Saúde

CROSS – Central de Regulação de Ofertas de Serviços de Saúde

GLASS – *Global Antimicrobial Resistance Surveillance System*

IPRS – Índice Paulista de Responsabilidade Social

IRAS – Infecções Relacionadas à Assistência em Saúde

KPC – *Klebsiella pneumoniae carbapenemases*

LightGBM – *Light Gradient Boosting Machine*

LISA – *Anselin Local Moran's I*

ML – *machine learning*

MRSA – *Staphylococcus aureus* resistente à meticilina

MSE – Erro quadrático médio

OMS – Organização Mundial da Saúde

ONU – Organização das Nações Unidas

PCA – Análise de Componentes Principais

ROC – *Receiver Operator Characteristic*

RRAS – Redes Regionais de Atenção à Saúde

SUS – Sistema Único de Saúde

UTI – Unidade de Terapia Intensiva

VN – Verdadeiro Negativo

VP – Verdadeiro Positivo

XGBoost – *Extreme Gradient Boosting*

SUMÁRIO

1. INTRODUÇÃO

As Infecções Relacionadas à Assistência à Saúde (IRAS), anteriormente denominadas infecções hospitalares, são aquelas adquiridas durante o cuidado prestado em qualquer ambiente de saúde, incluindo hospitais, ambulatórios, clínicas, instituições de longa permanência, serviços de hemodiálise, atendimento domiciliar e consultórios odontológicos. A substituição do termo “infecção hospitalar” por IRAS reflete a compreensão ampliada de que o risco infeccioso não está restrito ao ambiente hospitalar, podendo ocorrer em toda a rede de atenção à saúde (CENTRO DE VIGILÂNCIA EPIDEMIOLÓGICA, 2012). Essas infecções representam um importante desafio à saúde pública mundial, contribuindo para o aumento da morbimortalidade, da resistência bacteriana e dos custos assistenciais. Em 2017, a Organização Mundial da Saúde (OMS) publicou uma lista global de bactérias prioritárias que necessitam urgentemente de novos antibióticos, destacando patógenos associados às IRAS, como as enterobactérias resistentes a múltiplas drogas (WHO, 2017). No Brasil, a ANVISA reforça que o uso inadequado de antibióticos, comum em ambientes de atenção à saúde, favorece o surgimento de bactérias resistentes e compromete o controle efetivo das IRAS (ANVISA, 2022).

A resistência bacteriana aos antibióticos é uma preocupação crescente da saúde pública global, comprometendo a eficácia terapêutica e agravando os desfechos clínicos de infecções comuns. Estudos estimam que aproximadamente 63,5% das infecções causadas por bactérias resistentes a antibióticos estão associadas à prestação de cuidados em saúde. Esse quadro tem implicações diretas na mortalidade global, uma vez que, em 2019, mais de 5 milhões de óbitos foram relacionados à resistência antimicrobiana, evidenciando a gravidade e a escala do problema em nível mundial (WHO, 2023). A resistência pode surgir tanto de forma natural, por mecanismos intrínsecos, quanto ser adquirida por transferência horizontal de genes ou mutações genéticas (BUSH, 2020). Como resposta a esse desafio, alguns países implementaram estratégias coordenadas para conter a disseminação dessas bactérias nos ambientes hospitalares. Um exemplo foi a intervenção nacional realizada em Israel, que demonstrou eficácia na redução da propagação de enterobactérias resistentes aos carbapenêmicos em hospitais de cuidados pós-

agudos, obtendo sucesso na redução da transmissão nosocomial[1] (BEN-DAVID *et al.*, 2012). *Acinetobacter ssp* e enterobactérias são uma ameaça sem precedentes à saúde humana (LOGAN; WEINSTEIN, 2017). Esses organismos são um patógeno prioritário para a OMS (WHO, 2017) e os Centros de Controle e Prevenção de Doenças (Centers for Disease Control and Prevention, 2013)

Desde o início dos anos 2000, o surgimento de cepas de enterobactérias produtoras de *Klebsiella pneumoniae carbapenemase* (KPC) tem sido relatado nos Estados Unidos e em todo o mundo (NORDMANN; CUZON; NAAS, 2009). A KPC é uma bactéria multirresistente capaz de disseminar-se em ambientes hospitalares, desafiando os tratamentos convencionais devido à sua resistência à classe de antibióticos carbapenêmicos, considerados como último recurso para infecções graves (WHO, 2023). Reconhecendo sua ameaça à saúde pública, a KPC foi classificada como uma ameaça prioritária pelos Centros de Controle e Prevenção de Doenças dos Estados Unidos e pela Organização Mundial da Saúde (WHO, 2017; BERRÍOS-PASTÉN *et al.*, 2020; WHO, 2023). O conhecimento da disseminação da KPC entre os hospitais é importante para o desenvolvimento e implementação de estratégias eficazes de controle e prevenção.

A capacidade do Acinetobacter baumannii de adquirir resistência a múltiplos antibióticos, especialmente aos carbapenêmicos, contribui significativamente para sua virulência e para a dificuldade no tratamento das infecções que provoca (ANVISA, 2021; MOUBARECK; HALAT, 2020). Embora estudos nacionais indiquem prevalências elevadas, com taxas de resistência a carbapenêmicos superiores a 70% em pacientes internados em Unidades de Terapia Intensiva (ANVISA, 2021), os dados da presente pesquisa revelam uma tendência de redução dessas infecções ao longo do tempo. No entanto, observou-se uma ampliação do seu espalhamento geográfico, o que indica a necessidade de abordagens integradas que combinem vigilância epidemiológica, análise espacial e modelos preditivos para o controle eficaz desse patógeno.

As enterobactérias, incluindo Escherichia coli e Klebsiella spp., destacam-se por sua capacidade de desenvolver resistência a antibióticos. Essas cepas são frequentemente associadas à produção de carbapenemases, e à aquisição de genes de resistência a cefalosporinas de última geração, fluoroquinolonas e outros

[1] A transmissão nosocomial refere-se à disseminação de microrganismos infecciosos dentro de um ambiente hospitalar, geralmente entre pacientes, profissionais de saúde ou visitantes.

antibióticos amplamente utilizados na prática clínica (KITABA *et al.*, 2024; WHO, 2019; NORDMANN *et al.*, 2012; PATERSON, 2006).

Segundo o relatório global mais recente da Organização Mundial da Saúde, a disseminação de bactérias multirresistentes, como as enterobactérias resistentes aos carbapenêmicos, permanece uma ameaça crítica à segurança do paciente, especialmente em unidades de terapia intensiva e ambientes hospitalares com alta complexidade assistencial (WHO, 2024). Nesse cenário, estudos de modelagem matemática e computacional têm sido desenvolvidos para compreender e predizer a dinâmica de proliferação dessas bactérias resistentes, com o objetivo de subsidiar estratégias de controle e prevenção mais eficazes.

Com o objetivo de compreender formas de reduzir a propagação de infecções hospitalares no sistema de saúde francês, incluindo organismos multirresistentes, Nekkab *et al.* (2017) realizaram um estudo que envolveu a construção de 1.000 redes aleatórias de pacientes a partir de uma rede geral. Para identificar padrões de agrupamento e dinâmicas de transmissão, os pesquisadores utilizaram o algoritmo Greedy, conhecido pela sua eficácia na maximização da modularidade em redes complexas, e o algoritmo "Equação do Mapa", desenvolvida por Rosvall e Bergstrom (2008), que se baseia na minimização do comprimento da descrição de fluxos de informação para detectar comunidades em redes. O estudo utilizou dados de altas hospitalares coletados ao longo de 2014, fornecendo informações importantes sobre o impacto da mobilidade dos pacientes na disseminação de patógenos dentro dos hospitais.

Uma análise distinta de três grupos de pacientes foi realizada: os infectados, os suspeitos de infecção e os pacientes em geral (rede completa). Os resultados mostraram que as movimentações dos pacientes são heterogêneas e centralizadas, indicando que, embora as movimentações variem em intensidade, a rede geral tem maior potencial para propagar infecções, considerando sua estrutura de centralidade e conectividade. Essa descoberta enfatiza a importância de estratégias de controle mais específicas e dirigidas, considerando a mobilidade dos pacientes como um fator primordial na dinâmica de transmissão de infecções hospitalares. Essa análise reforça a necessidade de intervenções de controle de infecção que abordem não apenas os pacientes já diagnosticados, mas também os movimentos de pacientes dentro do hospital, como medida para mitigar o risco de infecções nosocomiais, especialmente

aquelas causadas por microrganismos multirresistentes, que são particularmente difíceis de tratar (PITTET *et al.*, 2009; TACCONELLI *et al.*, 2018).

A relação entre o grau de conexão hospitalar e as taxas de infecções causadas pelo patógeno Staphylococcus aureus resistente à meticilina (MRSA), adquiridas em hospitais da Inglaterra e Holanda, foi estudada por Donker *et al.* (2012). Para testar essa associação, os pesquisadores mapearam o movimento de pacientes com base nas estatísticas de episódios hospitalares. A rede de contato dos pacientes internados em hospitais foi estruturada considerando que, primeiramente, os pacientes têm maior probabilidade de interagir com outros pacientes na mesma enfermaria, em seguida, eles tendem a ter contato com pacientes em hospitais do mesmo agrupamento hospitalar. Essa estrutura agrava a dispersão de infecções hospitalares e influencia na rápida velocidade com que as infecções se espalham (DONKER *et al.*, 2012).

Ao comparar o padrão de referência nos dois países, os hospitais ingleses apresentam uma disseminação mais rápida das infecções (DONKER *et al.*, 2012). Os resultados mostraram também que os hospitais não podiam ser considerados unidades individuais, mas sim elementos interconectados de redes modulares maiores. Dessa forma, a disseminação das infecções influencia o planejamento dos sistemas de saúde, a gestão de pacientes e o controle das infecções hospitalares.

Uma modelagem matemática juntamente com análise de redes que sustentam a hipótese de que os movimentos de pacientes entre hospitais alemães afetam a transmissão nosocomial, no âmbito regional e nacional (Ciccolini *et al.*, 2013). Os autores encontraram que é possível a realização de planejamento do controle de infecção hospitalar local, regional, nacional e transfronteiriço do patógeno bacteriano nosocomial MRSA. A disseminação de novos patógenos nosocomiais MRSA foi estudada com a utilização de um modelo de epidemia de infecção suscetível hospitalar (CICCOLINI *et al.*, 2014). Cada hospital pode estar em um dos dois estados possíveis: suscetível e livre do patógeno ou afetado pelo patógeno. Cada paciente que recebe alta de um hospital afetado está associado a uma probabilidade de disseminar o patógeno no próximo hospital de admissão (CICCOLINI *et al.*, 2014).

Assim, uma vez que um hospital se torne infectado, os pacientes que recebem alta e são posteriormente admitidos em outras instituições, o que ocorre em aproximadamente 50% dos casos dentro de um intervalo de 17 a 25 dias, apresentam

uma probabilidade constante, embora baixa, de introduzir o patógeno no novo hospital de destino.

No Brasil foi analisado os determinantes socioeconômicos do consumo de antibióticos no estado de São Paulo entre os anos de 2008 e 2012, com ênfase no impacto da proibição da venda de antibióticos sem prescrição médica, implementada em outubro de 2010 (KLIEMANN *et al.* 2016). O estudo revelou que o consumo de antibióticos aumentou até 2010, atingindo 9,95 doses diárias definidas por 1.000 habitantes por dia (DID), e sofreu uma redução significativa após a intervenção, caindo para 8,06 DID em 2012. A análise demonstrou que municípios com maior proporção de população urbana, maior densidade de estabelecimentos privados de saúde e menor nível de analfabetismo apresentaram maiores níveis de consumo. Além disso, observou-se que o efeito da política foi mais pronunciado em municípios com maior proporção de mulheres. Esses resultados indicam que fatores demográficos e estruturais influenciam tanto os níveis de consumo quanto a efetividade das políticas de restrição de vendas, ressaltando a importância de considerar as disparidades socioeconômicas regionais na formulação de estratégias de controle.

A descrição e identificação do comportamento espaço-temporal dos bacilos Gram-negativos, entre eles *Acinetobacter baumannii* e enterobactérias resistentes aos carbapenêmicos, foi realizada para os municípios brasileiros e foi mais detalhada para os municípios do estado de São Paulo nos anos de 2011 a 2016 (CARAZATTO, 2019). Os resultados apontam para uma concentração inicial dos relatos de CRAB e CRE no Sudeste Brasileiro, seguida de uma dispersão pelo país. Constatou-se que microrganismos multidroga-resistentes emergem em áreas mais populosas, sofrendo disseminação centrífuga a partir da capital do estado de São Paulo (CARAZATTO, 2019, p. 87).

A Organização Mundial da Saúde (OMS) classificou as enterobactérias resistentes a múltiplos fármacos como patógenos críticos na sua lista global de prioridades para pesquisa e desenvolvimento de novos antibióticos, destacando a urgência em desenvolver estratégias eficazes para sua prevenção e controle (WHO, 2017). Essa priorização reflete a crescente ameaça que essas bactérias representam à saúde pública mundial, especialmente em ambientes hospitalares, devido à sua elevada capacidade de disseminação e à limitação de opções terapêuticas disponíveis (TACCONELLI *et al.*, 2018).

Um estudo tratou da simulação do movimento de pacientes dentro e entre hospitais para avaliar a dinâmica de transmissão e a incidência de enterobactérias produtoras de KPC (Vilches *et al.*, 2019). Os autores desenvolveram um modelo de metapopulação onde as conexões entre hospitais são feitas por meio de um modelo teórico de rede hospitalar, com base nos tamanhos e nas distâncias entre os hospitais brasileiros. Utilizando um sistema de equações diferenciais ordinárias (modelo multipatch), os autores avaliaram o impacto da movimentação de pacientes na dinâmica da transmissão. Os resultados indicaram que as transferências inter-hospitalares são um fator crítico para a propagação regional dos patógenos multirresistentes e que a ausência de controle local pode amplificar rapidamente surtos em escala sistêmica. O estudo destaca a necessidade de políticas coordenadas de controle de infecção que integrem dados de rede e o rastreamento do fluxo de pacientes entre unidades hospitalares.

Entre 2017 e 2021, um estudo realizado na Etiópia identificou altas taxas de resistência em *Escherichia coli* e *Klebsiella pneumoniae*, com destaque para o aumento significativo da resistência à ciprofloxacina (90%) e aos carbapenêmicos (38%) em *K. pneumoniae*, evidenciando a urgência de medidas de controle e vigilância, sobretudo em contextos com infraestrutura laboratorial limitada (KITABA *et al.*, 2024).

A mortalidade entre pacientes infectados por *Acinetobacter spp.* está fortemente associada aos cuidados de saúde, com destaque para a pneumonia relacionada ao uso de ventilação mecânica e para as infecções da corrente sanguínea, que se apresentam como as mais prevalentes. Essas infecções podem variar em frequência, de 5% em enfermarias gerais a até 54% em unidades de terapia intensiva, conforme evidenciado em um estudo realizado no Hospital Universitário de Toronto, Canadá, envolvendo 1.100 leitos, considerando o período de 1985 a 1995 (POUTANEN; LOUIE; SIMOR, 1997). Uma revisão de escopo publicada recentemente reforça a importância da vigilância da resistência antimicrobiana (RAM) em ambientes de atenção primária à saúde, destacando a subutilização de dados locais na orientação de decisões clínicas e estratégias de *stewardship*. A análise revelou que mais de 80% das prescrições de antibióticos ocorrem fora do ambiente hospitalar, muitas vezes sem suporte por sistemas de vigilância em tempo real. O estudo aponta que a ausência de plataformas digitais integradas, somada à fragmentação dos dados e à falta de capacitação dos profissionais, limita a eficácia das ações de controle da

RAM nesse nível de atenção. Nesse contexto, os autores defendem a padronização e digitalização dos sistemas de vigilância, com acesso facilitado e visualização amigável dos dados, como medidas essenciais para ampliar o uso de evidências locais na prescrição racional de antimicrobianos e no planejamento de políticas públicas (MORI *et al.*, 2025). Esse grupo de bactérias integra a lista de patógenos prioritários publicada pela Organização Mundial da Saúde em 2017, que destaca os microrganismos mais preocupantes para a saúde pública devido à resistência a múltiplos antimicrobianos. Na categoria de prioridade 1 – crítica, estão incluídas as enterobactérias resistentes aos carbapenêmicos, como *Klebsiella pneumoniae*, *Escherichia coli*, *Enterobacter spp.*, além de *Acinetobacter baumannii*, todas associadas a altas taxas de mortalidade e a infecções graves em ambientes hospitalares (WHO, 2017).

Os fatores espaciais e sociodemográficos associados às taxas de infecção do sítio cirúrgico em hospitais do interior do Estado de São Paulo foram analisados por Carvalho *et al.* (2021), por meio de modelos estatísticos univariados e multivariados de regressão binomial inflada de zeros. Os resultados evidenciaram que a incidência às taxas de infecção foi significativamente associada à distância em relação à capital, com maiores taxas observadas em hospitais mais afastados, especialmente nas regiões centrais e no extremo oeste do estado. A análise espacial por meio de mapas de densidade de kernel permitiu identificar áreas de maior concentração de infecções. Apesar de variáveis estruturais como o porte hospitalar e a presença de UTIs não apresentarem associação significativa após o ajuste multivariado, os autores sugerem que fatores como desigualdades socioeconômicas, deficiências estruturais e escassez de recursos humanos e técnicos nos hospitais públicos mais periféricos podem contribuir para os elevados índices das às taxas de infecção em sítios cirúrgicos.

Uma investigação sobre os fatores de risco associados à prevalência da colonização/infecção por CRE nos pacientes durante sua permanência na UTI, em 24 unidades de terapia intensiva, na China, foi conduzido por Hu *et al.* (2023). Foram utilizados modelos de regressão logística multivariada na análise de associação. O estudo revelou que 8,5% dos pacientes adquiriram resistência aos carbapenêmicos durante a internação na UTI, com uma alta taxa de colonizações (67,6%), destacando a necessidade urgente de estratégias de controle mais eficazes para reduzir a disseminação desses patógenos nas UTIs.

De acordo com o relatório da ANVISA referente ao período de janeiro de 2015 a dezembro de 2016, foram registradas 2.129 infecções por *Acinetobacter baumannii*, das quais 945 (44,4%) foram causadas por cepas resistentes aos carbapenêmicos (CRAB). No mesmo intervalo, foram notificadas 4.997 infecções por enterobactérias, sendo 1.636 (32,7%) relacionadas a cepas resistentes a carbapenêmicos (CRE). Esses dados refletem o cenário da resistência microbiana em instituições de saúde brasileiras durante o biênio 2015–2016, com destaque para a elevada proporção de resistência em agentes clínicos críticos. Já segundo o Boletim Epidemiológico da ANVISA publicado em 2021, o complexo *Acinetobacter baumannii* foi o quarto patógeno mais frequente em infecções primárias de corrente sanguínea causadas por microrganismos resistentes aos carbapenêmicos. Em pacientes internados em Unidades de Terapia Intensiva, a taxa de resistência aos carbapenêmicos chegou a 79,5% nas infecções por esse patógeno no ano de 2019 (ANVISA, 2021).

Portanto, entender a dinâmica de disseminação espaço-temporal e os mecanismos de resistência dessas bactérias é importante para o desenvolvimento de estratégias eficazes de controle e prevenção de infecções nosocomiais. Estudos recentes têm explorado modelos de transmissão e fatores epidemiológicos associados à propagação desses patógenos, fornecendo uma nova compreensão para melhorar as práticas de controle de infecção hospitalar e a gestão de antimicrobianos. Esta abordagem combina conceitos da epidemiologia, microbiologia, cartografia e análise espacial, permitindo uma compreensão abrangente dos padrões de distribuição geográfica, as rotas de transmissão e os fatores de risco associados.

Nas últimas décadas, diversos pesquisadores têm proposto modelos preditivos aplicados à predição de doenças. Inicialmente, esses modelos eram baseados em análises multivariadas e regressão logística (HOSMER; LEMESHOW, 2000). Com o desenvolvimento da tecnologia, começaram a ser utilizados algoritmos de *machine learning* na modelagem preditiva em Saúde. Esses algoritmos são baseados em informações incluídas na representação dos dados, permitindo a análise de como essas características se correlacionam com a disseminação de bactérias resistentes a múltiplas drogas (KE *et al.*, 2017). O uso de *machine learning* na predição de doenças tem mostrado potencial significativo em melhorar a precisão dos diagnósticos e a eficiência dos tratamentos de saúde (ESTEVA *et al.*, 2017).

A predição consiste na aplicação de algoritmos de *machine learning* para compreender a estrutura dos dados existentes e gerar regras de predição. Na área da

saúde, modelos preditivos podem ser empregados para estimar o risco de determinado desfecho ocorrer, dado um conjunto de características socioeconômicas, demográficas, relacionadas ao hábito de vida e às condições de saúde, entre outras (SANTOS *et al.*, 2019). Alguns algoritmos de aprendizagem de máquina mais conhecidos incluem as Redes Neurais Artificiais, Floretas Aleatórias, P*erceptron*, *Support Vector Machine, Gradiente Boosting Classifier*, *Extreme Gradient Boosting* (XGBoost), *Categorical Boosting* (CatBoost), *Light Gradient Boosting Machine* (LightGBM).

As Redes Neurais Convolucionais (CNNs) têm se destacado principalmente no diagnóstico por imagem. Essas redes são altamente eficazes no reconhecimento de padrões em imagens médicas, como radiografias, tomografias computadorizadas (CT) e ressonâncias magnéticas (MRI). Um estudo realizado por Chan *et al.* (2020) empregou as CNNs para a detecção de câncer de mama. Os resultados mostraram que a CNN apresentou uma área sob a curva (AUC – *Area Under the Curve*) significativa, indicando que o modelo foi altamente eficaz em diferenciar mamografias malignas de benignas em estágio inicial.

Entre as aplicações recentes de algoritmos de *machine learning* na saúde pública brasileira, destaca-se um estudo ecológico que estimou taxas de mortalidade por câncer em municípios do país, com o intuito de identificar aglomerados espaciais estatisticamente significativos de mortalidade excessiva não explicada por características sociodemográficas locais (TEIXEIRA *et al.*, 2023). O modelo de *gradient boosting* (XGB) apresentou o melhor desempenho preditivo, com coeficiente de determinação $R^2 = 0,66$, valor superior ao obtido por outros algoritmos testados, como Floresta Aleatória ($R^2 = 0,65$), *Support Vector Machine* ($R^2 = 0,60$) e *Least Absolute Shrinkage and Selection Operator* (LASSO) com $R^2 = 0,59$. A metodologia permitiu localizar regiões prioritárias para investigação epidemiológica, como Bagé (RS), Porto Velho (RO) e Macapá (AP), onde foram observadas taxas de mortalidade por câncer superiores ao esperado. As variáveis mais relevantes para a predição incluíram a proporção de população branca e o acesso a computadores, refletindo a influência de determinantes sociais na distribuição espacial dos óbitos por câncer no Brasil.

Utilizando dados do Sistema de Informação de Agravos de Notificação (SINAN), base nacional mantida pelo Ministério da Saúde do Brasil, um estudo recente aplicou algoritmos de machine learning para predizer a ocorrência de perda de

seguimento durante o tratamento da tuberculose, conhecida como *Loss to Follow-Up* (LTFU). A base analisada incluiu 243.726 casos notificados entre 2015 e 2022, dos quais 41.373 foram classificados como LTFU e 202.353 como tratamentos concluídos com sucesso. Foram testados modelos de Regressão Logística, Floresta Aleatória e LightGBM, sendo este último o que apresentou o melhor desempenho preditivo. A escolha do modelo considerou, principalmente, a área sob a curva ROC (AUC), que variou entre 0,71 e 0,72 (RODRIGUES *et al.*, 2024).

No cenário internacional, diversos modelos foram propostos, mas não foi encontrado nenhum para predição e proliferação espacial da transmissão das bactérias resistentes a multidrogas, em especial a CRE e a CRAB. As abordagens tradicionais de monitoramento da resistência microbiana, baseadas em métodos estatísticos convencionais, apresentam limitações em sua capacidade de capturar a complexa dinâmica da proliferação espacial e temporal das bactérias resistentes. O *machine learning* surge como uma ferramenta para superar essas limitações, pois podem aprender mecanismos de disseminação da bactéria a partir de um conjunto de variáveis, tais como as infecções sanguíneas, bem como as estruturas hospitalares e dados socioeconômicos de cada região.

A interpretação dos resultados de modelos de *machine learning* treinados tem se mostrado uma abordagem promissora tanto para validar a acurácia dos modelos quanto para explorar padrões associados aos mecanismos de disseminação da resistência antimicrobiana (YOO *et al.*, 2024). Essa modelagem possibilita o desenvolvimento de estratégias mais eficazes de controle e prevenção da resistência, sobretudo em infecções causadas por patógenos críticos como *Acinetobacter baumannii* e enterobactérias produtoras de KPC (RODRIGUES *et al.*, 2024). Recentemente, avanços foram obtidos por meio da aplicação de técnicas como *feature extraction* e algoritmos de aprendizado supervisionado, como LightGBM e Floresta Aleatória, com acurácia superior a 90% na predição de resistência a múltiplos antibióticos, com base em sequências genômicas (RASTOGI *et al.*, 2024). Apesar desses progressos, a aplicação prática de modelos preditivos no contexto da saúde pública brasileira ainda é limitada.

Os algoritmos de *machine learning*, em especial as Florestas Aleatórias, representam abordagens promissoras para predição com base em conjuntos de dados tabulares. Além de sua flexibilidade para tarefas de regressão e classificação, ou seja, para predição de características quantitativas e categóricas, esses modelos

oferecem medidas de importância das variáveis, permitindo classificar os preditores conforme sua relevância preditiva (DEGENHARDT; SEIFERT; SZYMCZAK, 2019). Nesse contexto, uma análise comparativa entre treze algoritmos de *machine learning* amplamente utilizados, com exceção das Redes Neurais, foi conduzida a partir de 165 bases de dados distintas, com o objetivo de fornecer recomendações baseadas em evidências para orientar pesquisadores na escolha de métodos mais adequados a diferentes tipos de problemas (OLSON *et al.*, 2018).Os resultados da análise permitiram a recomendação de cinco algoritmos, com hiperparâmetros que maximizam o desempenho do classificador nos problemas testados, bem como orientações gerais para aplicação de aprendizado de máquina em problemas de classificação supervisionada, sendo eles: *Gradiente Boosting Classifier*, Floresta Aleatória, *Support Vector Classifier*, *Extra Tree*, *Stochastic Gradient Descent*, (Figura 1).

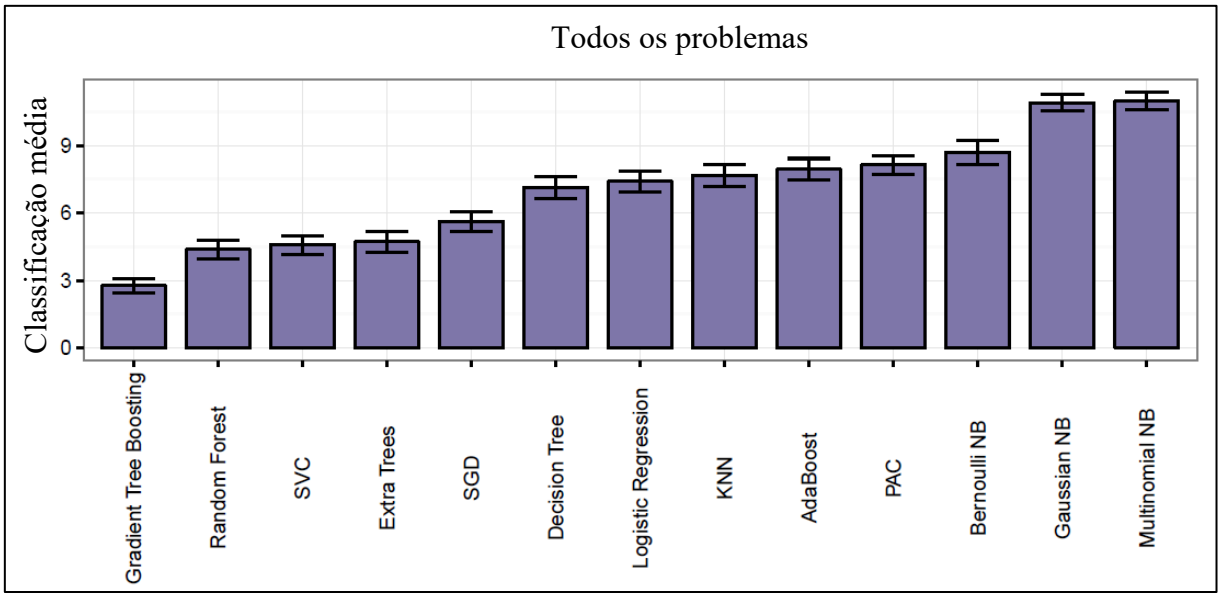

Figura 1 – Classificação média dos algoritmos de ML em todos os conjuntos de dados. As barras de erro indicam o intervalo de confiança de 95% para o modelo de Classificação.
Fonte: Adaptado de Olson *et al.* (2018)

Outros trabalhos utilizando algoritmos baseados em árvores, como os de árvores de decisão e *gradient boosting*, frequentemente superam outros algoritmos de *machine learning* em termos de capacidade preditiva, oferecendo resultados superiores na maioria dos cenários de análise de dados (RASCHKA e MIRJALILI, 2017; GÉRON, 2017).

Esta pesquisa busca encontrar modelos preditivos e variáveis associadas na identificação de padrões de disseminação de bactérias multirresistentes, como *Acinetobacter spp*. e enterobactérias, em hospitais e municípios do estado de São Paulo. A análise preditiva emprega o seguinte conjunto de algoritmos: Florestas Aleatórias, Redes Neurais, XGBoost, LightGBM, Catboost e Florestas Aleatórias.

A hipótese desta pesquisa é que fatores socioeconômicos em conjunto com variáveis hospitalares, que refletem a qualidade da assistência à saúde nos hospitais ou municípios, têm influência significativa na análise preditiva que visa encontrar explicações para a disseminação de infecções causadas por bactérias resistentes a múltiplas drogas. Este trabalho de pesquisa pretende responder às seguintes questões:

1. Como pode ser explicada a tendência temporal (valores absolutos e taxas) da disseminação de bactérias multirresistentes no estado de São Paulo?
2. Qual é o padrão de distribuição espacial das taxas de *Acinetobacter spp.* e das taxas de enterobactérias por município no estado de São Paulo, ao longo do tempo?
3. Existe associação espacial entre fatores socioeconômicos e a distância do município até a capital com as taxas de *Acinetobacter spp.* ou taxas de enterobactérias no estado de São Paulo, ao longo do tempo?
4. Qual algoritmo de *machine learning* e variáveis explicativas melhor prediz a disseminação de bactérias multirresistentes no estado de São Paulo?

1.1. Objetivos

Esta pesquisa tem como objetivo investigar a disseminação de bactérias multirresistentes e seus fatores associados no Estado de São Paulo, entre 2011 e 2019, utilizando análise de padrão de distribuição espacial e modelagem preditiva baseada em algoritmos de *machine learning*. Os objetivos específicos compreendem:

- Verificar o comportamento da tendência temporal da disseminação de *Acinetobacter spp.* e enterobactérias no estado de São Paulo;
- Descrever e identificar o padrão de distribuição espaço-temporal das taxas de *Acinetobacter spp.* e das taxas de enterobactérias nos município do estado de São Paulo;

- Examinar a associação entre as taxas de *Acinetobacter spp.* e taxas de enterobactérias com fatores socioeconômicos nos municípios do estado de São Paulo;
- Analisar o desempenho de diferentes algoritmos de predição para a disseminação de *Acinetobacter spp.* e enterobactérias nos hospitais e municípios do estado de São Paulo.
- Determinar, a partir da modelagem preditiva, a contribuição das variáveis associadas na disseminação de bactérias multirresistentes.

1.2. Justificativa

O surgimento e disseminação de bactérias hospitalares resistentes a múltiplas drogas representam um desafio significativo para a saúde pública global pelo fato de as bactérias desenvolverem mecanismos complexos de resistência que as tornam refratárias aos tratamentos convencionais com antibióticos. Este fenômeno não só aumenta a morbidade[2] e mortalidade[3] associadas a infecções hospitalares, mas também amplia os custos de saúde e compromete a eficácia dos sistemas de saúde em todo o mundo.

Ao longo das últimas décadas, pesquisadores têm se dedicado a estudar a resistência antimicrobiana e suas implicações na esfera hospitalar. A análise espaço-temporal dessas bactérias resistentes torna-se fundamental para compreender padrões de disseminação, identificar áreas geográficas de maior vulnerabilidade e desenvolver estratégias de intervenção direcionadas. Este estudo busca explorar o problema das bactérias hospitalares resistentes a múltiplas drogas, considerando as dimensões espacial e temporal, a fim de oferecer uma compreensão abrangente e aprofundada desse fenômeno.

[2] MORBIDADE refere-se ao conjunto dos indivíduos que adquirem doenças (ou determinadas doenças) num dado intervalo de tempo em uma determinada população. A morbidade mostra o comportamento das doenças e dos agravos à saúde na população.

[3] MORTALIDADE: refere-se ao conjunto dos indivíduos que morreram num dado intervalo do tempo. Representa o risco ou probabilidade que qualquer pessoa na população apresenta de poder vir a morrer ou de morrer em decorrência de uma determinada doença (MINISTÉRIO DA SAÚDE, 2009).

Em 2015, a Organização Mundial de Saúde (OMS) desenvolveu um plano global contra a crescente ameaça da resistência aos antimicrobianos conhecido como *Global Antimicrobial Resistance Surveillance System* – GLASS (WHO, 2021), uma plataforma para o compartilhamento de dados sobre resistência aos antimicrobianos no mundo. No ano de 2017, o Ministério da Saúde brasileiro aderiu ao sistema e, em 2018, iniciou o próprio programa nacional de vigilância antimicrobiana (BR-GLASS), sendo seu estudo piloto iniciado no estado do Paraná. No âmbito do BR-GLASS (*Global Antimicrobial Resistance Surveillance System*), Pillonetto *et al.* (2020), enfatizou que é necessária uma preocupação com as altas taxas de resistência a carbapenêmicos em *Acinetobacter spp.*, que atingiram 81,4% em 2018. Esse trabalho realizado no estado do Paraná mostrou a relevância do monitoramento contínuo e da implementação de políticas rigorosas de controle de infecções para mitigar a propagação desses patógenos. O estudo também enfatizou a importância da colaboração internacional para entender e combater a resistência antimicrobiana em um contexto global.

Assim, os resultados desta tese podem contribuir para a formulação de políticas públicas que são direcionadas ao controle e prevenção dos patógenos supracitados, otimizando a alocação de recursos e promovendo a saúde pública. Espera-se, também, que a análise espacial revele padrões e tendências na proliferação dessas bactérias, permitindo o desenvolvimento de estratégias de controle mais eficazes e direcionadas.

As infecções por bactérias resistentes a multidrogas aumentam a morbimortalidade do paciente e impõem custos de saúde (ZIMLICHMAN *et al.*, 2013), além da permanência hospitalar e ações jurídicas contra o hospital, para indenizações. O indivíduo portador desta bactéria geralmente não apresenta sintomas, sendo que a sua presença é identificada quando o tratamento recomendado não mostra eficácia.

A resistência antimicrobiana tornou-se uma ameaça global à saúde pública, que pode levar a até dez milhões de mortes por ano até 2050 (O'NEILL, 2016, p. 1; HARRIS *et al.*, 2023), e um acumulado de 100 trilhões de dólares da produção econômica estão em risco devido ao aumento de resistência a medicamentos. A infecção do sítio cirúrgico, por exemplo, representa uma ameaça particular em casos de países de renda média e baixa (CARVALHO, 2021). De acordo com o relatório da ONU, da Organização Mundial da Saúde e da Organização Mundial da Saúde Animal,

de 2019, as infecções resistentes aos medicamentos podem causar até 10 milhões de mortes anuais até 2050, além de levar até 24 milhões de pessoas à pobreza extrema até 2030, se nenhuma ação for tomada (WHO, 2019).

Os pacientes internados em unidades de terapia intensiva (UTI) apresentam uma alta suscetibilidade a colonizações e infecções hospitalares por organismos multirresistentes, incluindo o *Acinetobacter spp.* e a enterobactérias. Essa vulnerabilidade é atribuída a diversos fatores, incluindo comorbidades, procedimentos invasivos, uso de dispositivos médicos, fragilidade, internação prolongada e uso frequente de ANTIBIÓTICOS (YI; KIM, 2021 e BLOT *et al.*, 2022).

Já para uma análise espacial, a compreensão profunda e dinâmica da proliferação espaço-temporal das bactérias em diferentes regiões do estado de São Paulo é fundamental para o combate eficaz à resistência microbiana. O conhecimento dos padrões e tendências da proliferação permitirá a identificação de áreas de risco, o direcionamento de ações de vigilância e controle mais assertivas e a tomada de decisões estratégicas para conter a disseminação dessas bactérias.

De acordo com Sorre (1978, p. 238), é importante localizar os fenômenos, determinar as áreas atingidas por esses fenômenos, e marcar as variações de intensidade dentro dessas áreas. A análise espacial constitui uma ligação entre a cartografia e as áreas de análise aplicada. De tal modo, é possível realizar análises espaciais que ajudam a compreender a complexidade do problema, associando a outras variáveis, com dados sociais e econômicos, com os quais permitem uma discussão mais aprofundada e atualizada da problemática, por exemplo, que possam agravar ou não o evento, e, também, que possam avaliar tendências futuras.

De acordo com o boletim epidemiológico da Agência Nacional de Vigilância Sanitária (ANVISA), do ano de 2021, dentre os microrganismos resistentes aos carbapenêmicos em infecção primária de corrente sanguínea, o complexo *Acinetobacter* baumannii foi o quarto patógeno mais frequente, apresentando 79,5% de resistência aos carbapenêmicos em pacientes com infecção de corrente sanguínea internados em Unidades de Terapia Intensiva (UTI), no ano de 2019 (ANVISA, 2021).

Assim, no mesmo ano, 2021, a ANVISA instituiu o Programa Nacional de Prevenção e Controle de Infecções Relacionadas à Assistência à Saúde para o período de 2021 a 2025, com o objetivo de reduzir em todo o país a incidência de infecções hospitalares, inclusive das bactérias resistentes à múltiplas drogas,

implementando práticas de prevenção e controle de infecções baseadas em evidências clínicas (ANVISA, 2021).

A morbidade associada às IRAS pode desencadear o uso de recursos hospitalares com maior duração. No trabalho de Kiffer *et al.* (2015), foi estabelecido modelo de custo de ocupação-dia em unidades hospitalares de 11 hospitais de grande porte no Brasil. O custo diário do paciente com IRAS foi 55% superior ao de um paciente sem IRAS.

Em 2017, a Organização Mundial da Saúde (OMS) publicou uma lista de patógenos prioritários resistentes a antibióticos. O grupo mais crítico inclui bactérias multirresistentes, como *Acinetobacter spp.* e Enterobacteriaceae, que podem causar infecções graves e muitas vezes fatais, como infecções da corrente sanguínea e pneumonia (WHO, 2017). A resistência aos antimicrobianos é uma preocupação global para a comunidade científica e para os sistemas de saúde, uma vez que bactérias resistentes são responsáveis por 700 mil mortes anuais, conforme dados da OMS (Jornal da USP, 2021; *Antimicrobial Resistance Collaborators*, 2022).

Nesse contexto, a detecção rápida e precisa das bactérias resistentes é importante para direcionar o tratamento clínico e orientar intervenções de prevenção e controle (TACCONELLI *et al.*, 2023). Bem como a análise espacial é fundamental para entender a distribuição geográfica e rotas de transmissão de bactérias resistentes, auxiliando na criação de políticas públicas mais eficazes para controle e prevenção de infecções nosocomiais.

4. CONSIDERAÇÕES FINAIS

A resistência antimicrobiana representa um dos maiores desafios de saúde pública da atualidade, com impactos diretos na morbimortalidade hospitalar, aumento de custos assistenciais e redução da eficácia terapêutica. Nesta perspectiva, a presente tese desenvolveu, de maneira integrada, a dinâmica espacial e a predição de bactérias multirresistentes, abordando duas frentes principais: a análise espacial dos padrões de resistência no estado de São Paulo e a aplicação de algoritmos de *machine learning* para predição de *Acinetobacter spp.* e de enterobactérias, no estado de São Paulo, entre 2011 e 2019. Cada etapa contribuiu de forma singular para o entendimento do fenômeno.

O primeiro eixo da pesquisa, centrado na análise espacial, evidenciou padrões significativos na distribuição de bactérias multirresistentes. Identificaram-se aglomerados principalmente nos municípios que compõem o DRS da Grande São Paulo, associados a fatores como elevada densidade populacional, número de leitos de UTI, consumo anual de energia elétrica e indicadores socioeconômicos. A utilização de técnicas como o Índice de Moran Global bivariado, juntamente com mapas temáticos e de símbolos proporcionais, permitiu identificar as áreas de maior risco, de modo que possa subsidiar a priorização de intervenções públicas.

As análises espaciais, embora eficazes para revelar padrões espaciais, não foram capazes de quantificar com precisão o impacto individual de cada variável explicativa. Do mesmo modo, os modelos lineares tradicionais apresentam limitações diante da complexidade e não linearidade das interações entre variáveis, frequentemente presentes em fenômenos como a resistência bacteriana.

Assim, utilizou-se na segunda parte a modelagem preditiva com algoritmos de *machine learning*, que demonstraram alto desempenho na predição da resistência bacteriana. O XGBoost destacou-se em tarefas de classificação em nível hospitalar, enquanto a Floresta Aleatória apresentou maior eficácia na regressão em nível municipal. O CatBoost, por sua vez, foi eficaz em lidar com dados categóricos de alta dimensionalidade. As variáveis mais influentes nos modelos incluíram indicadores de fluxo hospitalar (paciente-dia), número de leitos de UTI, densidade demográfica e PIB, corroborando a literatura sobre os determinantes da resistência antimicrobiana.

A utilização das abordagens espacial e preditiva proporcionou uma compreensão aprofundada do fenômeno, permitindo não apenas identificar padrões

espaciais de disseminação, mas verificar como os fatores associados podem contribuir para a disseminação das bactérias multirresistentes. Verifica-se a importância de incorporar ferramentas de análise espacial na vigilância epidemiológica, e modelos preditivos desenvolvidos com *machine learning* para oferecer suporte à tomada de decisão, tanto em nível hospitalar quanto municipal.

Os resultados desta tese reforçam a necessidade de abordagens interdisciplinares e baseadas em dados para enfrentar o avanço da resistência antimicrobiana. As análises apresentadas oferecem suporte para a formulação de políticas públicas direcionadas, alocação eficiente de recursos e desenvolvimento de estratégias de controle adaptadas às realidades locais e regionais. A combinação de variáveis socioeconômicas e hospitalares na modelagem preditiva se mostrou eficaz para orientar ações de prevenção e controle, alinhando-se aos princípios da vigilância em saúde e da epidemiologia de precisão.

Apesar das contribuições alcançadas, a pesquisa enfrentou limitações importantes. A ausência de dados clínicos detalhados sobre práticas de antibioticoterapia, controle de infecções e perfil microbiológico reduziu a capacidade explicativa dos modelos com regressão em nível hospitalar. Além disso, a análise com médias municipais pode ter mascarado variações hospitalares significativas. Trabalhos futuros podem considerar a integração de dados genômicos, ambientais (como variáveis climáticas) e informações sobre consumo de antimicrobianos em diferentes setores, ampliando o escopo da análise para um entendimento mais sistêmico do problema.

Por fim, a presente tese contribui para o avanço metodológico e aplicado na vigilância e controle da resistência bacteriana. Os modelos desenvolvidos e as análises espaciais propostas oferecem recomendações práticas que podem auxiliar gestores hospitalares e formuladores de políticas públicas. A aplicação de algoritmos de *machine learning*, aliados à análise geoespacial, destaca-se como uma estratégia promissora para enfrentar os desafios da resistência antimicrobiana, promovendo avanços na saúde pública e na gestão hospitalar.

REFERÊNCIAS

ABDI, H.; WILLIAMS, L. J. Principal Component Analysis. Wiley Interdisciplinary Reviews: **Computational Statistics**, v.2, 2010. pp. 433-459.

AFSHINNEKOO, E.; MEYDAN, C.; CHOWDHURY, S.; JAROUDI, D.; BOYER, C.; BERNSTEIN, N.; MARITZ, J. M.; REEVES, D.; GANDARA, J.; CHHANGAWALA, S.; AHSANUDDIN, S.; SIMMONS, A.; NESSEL, T.; SUNDARESH, B.; PEREIRA, E.; JORGENSEN, E.; KOLOKOTRONIS, S.-O.; KIRCHBERGER, N.; GARCIA, I.; GANDARA, D.; DHANRAJ, S.; NAWRIN, T.; SALETORE, Y.; ALEXANDER, N.; VIJAY, P.; HÉNAFF, E. M.; ZUMBO, P.; WALSH, M.; O'MULLAN, G. D.; TIGHE, S.; DUDLEY, J. T.; DUNAIF, A.; ENNIS, S.; O'HALLORAN, E.; MAGALHAES, T. R.; BOONE, B.; JONES, A. L.; MUTH, T. R.; PAOLANTONIO, K. S.; ALTER, E.; SCHADT, E. E.; GARBARINO, J.; PRILL, R. J.; CARLTON, J. M.; LEVY, S.; MASON, C. E. Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. Cell Systems, **PubMed**, v. 1, n. 1, pp. 72-87, 2015. DOI: 10.1016/j.cels.2015.01.001.

AGÊNCIA BRASIL. **Uso inadequado de antibióticos aumenta resistência de bactérias.** Disponível em: <https://agenciabrasil.ebc.com.br/saude/noticia/2019-11/uso-inadequado-de-antibioticos-aumenta-resistencia-de-bacterias>. Acesso em: 10 de julho de 2024.

_____. **OMS atualiza lista de bactérias que ameaçam saúde humana**. 2024. Disponível em: < https://agenciabrasil.ebc.com.br/internacional/noticia/2024-05/oms-atualiza-lista-de-bacterias-que-ameacam-saude-humana >. Acesso em: 15 de janeiro de 2025.

ALCÂNTARA, E. MANTOVANI, J.; ROTTA, L.; PARK, E.; RODRIGUES, T.; CARVALHO, F. C.; SOUZA FILHO, C. R. Investigating spatiotemporal patterns of the COVID-19 in São Paulo State, Brazil. **Geospatial Health**. 2020.

ALI, Z. A.; ABDULJABBAR, Z. H.; TAHIR, H.; ALMUFTI, S. M.. Exploring the Power of eXtreme Gradient Boosting Algorithm in *Machine learning*: a Review. **Academic Journal of Nawroz University,** v. 12, n. 2, pp. 320-334, 2023. DOI: 10.25007/ajnu.v12n2a1612. Disponível em: https://www.researchgate.net/publication/371202498_Exploring_the_Power_of_eXtreme_Gradient_Boosting_Algorithm_in_Machine_Learning_a_Review. Acesso em: 19 de dezembro de 2024.

ALLEL, K.; GARCÍA, P.; LABARCA, J.; MUNITA, J. M.; RENDIC, M.; GRUPO COLABORATIVO DE RESISTENCIA BACTERIANA; UNDURRAGA, E. A. Socioeconomic factors associated with antimicrobial resistance of *Pseudomonas aeruginosa*, *Staphylococcus aureus*, and *Escherichia coli* in Chilean hospitals (2008–2017). **Antimicrobial Resistance & Infection Control**, v. 10, n. 1, p. 1-10, 2021. DOI: 10.1186/s13756-021-00981-9.

ALMEIDA, E. **Econometria Espacial Aplicada**. Editora Alínea. Campinas, SP, 2012. 498p.

ALVIM, A. L. S., COUTO, B. R. G. M.; GAZZINELLI, Factores de riesgo para Infecciones relacionadas con la Asistencia Sanitaria causadas por Enterobacteriaceae productoras de Klebsiella pneumoniae carbapenemase: un estudio de caso control. **Enfermería Global**, Universidad de Murcia, n. 58, pp. 267-278. 2020. Disponível em: <https://doi.org/10.6018/eglobal.380951>. Acesso em: 10 de dezembro de 2024.

ANDERSON, T. W. **An introduction to multivariate statistical analysis**. 3rd ed. New York: John Wiley, 2003. 721p.

ANSELIN, L. Local Indicators of Spatial Association-LISA. **Geographical Analysis**. v. 27, n.2. 1995.

_____ . Global Spatial Autocorrelation – Bivariate, Differential and EB Rate Moran Scatter Plot. Geoda. 2020. Disponível em: <https://geodacenter.github.io/workbook/6b_local_adv/lab6b.html>. Acesso em: 04 de janeiro de 2022.

_____ . An Introduction to Spatial Data Science with GeoDa. Volume 1: Exploring Spatial Data. 2023. Disponível em: <https://lanselin.github.io/introbook_vol1/>. Acesso em: 15 de outubro de 2024.

ANSELIN, L.; SYABRI, I.; SMIRNOV, O. *Visualizing Multivariate Spatial Correlation with Dynamically Linked Windows*. New Tools for Spatial Data Analysis: Proceedings of the Specialist Meeting, Santa Barbara, 2002.

ANSELIN, L; REY, S. J. **Perspectives on spatial data analysis**. Heidelberg: Springer, Berlin. 2010.

ANVISA – Agência Nacional de Vigilância Sanitária. **Relatório sobre o ônus da infecção endêmica associada aos cuidados de saúde em todo o mundo**. Brasília, 2021. 61p.

_____. **Uso incorreto de antibiótico estimula superbactérias**. Ministério da Saúde. Publicado em 04/07/2022. Disponível em: <https://www.gov.br/anvisa/pt-br/assuntos/noticias-anvisa/2018/uso-incorreto-de-antibiotico-estimula-superbacterias>. Acesso em: 20 de novembro de 2024.

ANTIMICROBIAL RESISTANCE COLLABORATORS. **Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis**. Lancet 399 (10325), 629–655, 2022.

ARANGO-ARGOTY, G., GARNER, E., PRUDEN, A., HEATH, L. S., VIKESLAND, P. ZHANG, L. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. **Microbiome,** v. 6, n. 23 p. 1–15, 2019. Disponível em: <https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-018-0401-z>. Acesso em: 15 de março de 2025.

ASSEMBLEIA LEGISLATIVA DO ESTADO DE SÃO PAULO**. Índice Paulista de Responsabilidade Social – IPRS: banco de dados municipais.** Fundação Sistema

Estadual de Análise de Dados – SEADE. São Paulo. Disponível em: <https://iprs.seade.gov.br>. Acesso em: 11 de março de 2022.

AUCHINCLOSS, A. H.; GEBREAB, S. Y.; MAIR, C.; ROUX, A. V. D. A review of spatial methods in epidemiology, 2000– 2010. Annual Review of Public Health, v. 33, p. 107-122, 2012. DOI:10.1146/annurev-publhealth-031811-124655

BAILEY, T. C., GATRELL, A. C. **Interactive Spatial Data Analysis**. Longman, 1995. 432p.

BALASUBRAMANIAN, R.; VAN BOECKEL, T. P.; CARMELI, Y.; COSGROVE, S.; LAXMINARAYAN, R. Global incidence in hospital-associated infections resistant to antibiotics: An analysis of point prevalence surveys from 99 countries. **PLOS Medicine**, San Francisco, v. 20, n. 6, p. e1004178, 2023. Disponível em: <https://doi.org/10.1371/journal.pmed.1004178>. Acesso em: 07 de abril de 2025.

BARCELLOS, C.; SABROZA, P. C. Socio-environmental determinants of the leptospirosis outbreak of 1996 in western Rio de Janeiro: a geographical approach. **International Journal of Environmental Health Research**, v. 10, n. 4, p. 301–313, 2000. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/0960312002001500>. Acesso em: 13 de fevereiro de 2025.

BATISTA, A. F. M.; CHIAVEGATTO FILHO, A. P. *Machine learning* aplicado à Saúde. SBCAS, 2019. Disponível em: <http://www.midiacom.uff.br/sbcas2019/MachineLearning-Saude.pdf>. Acesso em: 02 de dezembro de 2021.

BALASUBRAMANIAN, R., Boeckel, T. P. V., CARMELI, Y., COSGROVE, S., Laxminarayan, R. Global incidence in hospital-associated infections resistant to antibiotics: An analysis of point prevalence surveys from 99 countries. **PLOS Medicine**. 2023. Disponível em: <https://doi.org/10.1371/journal.pmed.1004178>. Acesso em 20 de fevereiro de 2025. pp. 1-13.

BATISTA, G. E. A. P. A.; MONARD, M. C. An analysis of four missing data treatment methods for supervised learning. **Applied Artificial Intelligence**, v. 17, n. 5-6, 2010, pp. 519-533. Disponível em: <https://doi.org/10.1080/713827181>. Acesso em: 10 de janeiro de 2025.

BEN-DAVID, D. et al. Outcome of a national strategy to contain the spread of carbapenem-resistant Enterobacteriaceae in post-acute care facilities. **American Journal of Infection Control**, v. 42, n. 7, p. 704-709, 2014.

BERGSTRA, J.; BENGIO, Y. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, v. 13, p. 281–305, 2012. Disponível em: <https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf>. Acesso em: 11 de novembro de 2024.

BERRÍOS-PASTÉN, C.; ACEVEDO, R.; ARROS, P.; VARAS, M. A.; WYRES, K. L.; LAM, M. M. C.; HOLT, K. E.; LAGOS, R.; MARCOLETA, A. E. Properties of genes encoding transfer RNAs as integration sites for genomic islands and prophages in Klebsiella pneumoniae. **bioRxiv**. 2020. pp. 1-32.

BIBLIOTECA VIRTUAL EM SAÚDE. **Higienização das mãos na assistência à saúde**. Disponível em: <https://bvsms.saude.gov.br/higienizacao-das-maos-na-assistencia-a-saude/>. Acesso em: 05 de dezembro de 2022.

BILAL, H.; KHAN, M. N.; KHAN, S.; SHAFIQ, M.; FANG, W.; KHAN, R. U.; RAHMAN, M. U.; LI, X.; LV, Q.-L.; XU, B. (2025). The role of artificial intelligence and *machine learning* in predicting and combating antimicrobial resistance. **Computational and Structural Biotechnology Journal,** v. 27, pp. 423–439. Disponível em: <https://doi.org/10.1016/j.csbj.2025.01.006>. Acesso em: 04 de abril de 2025.

BISHOP, C. M. **Neural Networks for Pattern Recognition**. Oxford, 1995. 482 p.

BOSZCZOWSKI, I. **Análise espacial da ocorrência de infecções bacterianas da corrente sanguínea causadas por agentes multirresistentes em unidades de terapia intensiva do estado de São Paulo.** Tese de Doutorado. Programa de Pós-Graduação da Faculdade de Medicina da Universidade de São Paulo. 2016. 134 p.

BOUROCHE, J. M.; SAPORTA, G. Análise de dados. **Zahar Editores**. Rio de Janeiro, 1982.

BRADLEY, A. P. The use of the area under the roc curve in the evaluation of *machine learning* algorithms. **Pattern Recognit**. 1997. pp.1145-1159.

BRASIL. Ministério da Ciência, Tecnologia e Inovações. **Desenvolvido no LNCC, estudo inédito identifica bactérias resistentes em hospitais do Brasil.** 2016. Disponível em: <https://www.gov.br/lncc/pt-br/assuntos/noticias/ultimas-noticias-1/desenvolvido-no-lncc-estudo-inedito-identifica-bacterias-resistentes-em-hospitais-do-brasil>. Acesso em: 20 de agosto de 2024.

BRASIL. Ministério da Saúde. **Portaria nº 2.616, de 12 de maio de 1998**. Dispõe sobre diretrizes e normas para a prevenção e o controle das infecções hospitalares. Brasília. Disponível em: <https://portaldeboaspraticas.iff.fiocruz.br/biblioteca/portaria-no-2616-de-12-de-maio-de-1998>. Acesso em: 15 de agosto de 2021.

BREIMAN, L. Random Forests. *Machine learning*. **Springer.** 2001. pp. 5-32.

BROWNLEE, J. **XGBoost With Python: Gradient Boosted Trees With XGBoost and scikit-learn**. *Machine learning* Mastery, 2021. Edition v1.15. 108p.Disponível em: <https://books.google.com.br/books?hl=pt-PT&lr=&id=HgmqDwAAQBAJ&oi=fnd&pg=PP1&dq=Brownlee,+J.+(2019).+XGBoost+With+Python:+Gradient+Boosted+Trees+With+XGBoost+and+scikit-learn.+Machine+Learning+Mastery.&ots=nNfEk9QbPE&sig=zZsnWaj1j_rwxkcYI85JZmicbZU#v=onepage&q&f=false>. Acesso em: 10 de outubro de 2024.

BUSH, L. M. Considerações gerais sobre bactérias anaeróbias. Manual MSD – Versão saúde para a família. 2020. Disponível em: < https://www.msdmanuals.com/pt-br/casa/infec%C3%A7%C3%B5es/infec%C3%A7%C3%B5es-bacterianas-considera%C3%A7%C3%B5es-gerais/considera%C3%A7%C3%B5es-gerais-sobre-bact%C3%A9rias>. Acesso em 20 de dezembro de 2022.

CÂMARA, G; CARVALHO, M. S. **Análise Espacial de Dados Geográficos** – Análise de Eventos Pontuais. In DRUCK, S.; CARVALHO, M. S.; CÂMARA, G; MONTEIRO, A. M. V. Brasília: EMBRAPA, 2004. p.57-72.

CÂMARA, G; CARVALHO, M. S. CRUZ, O. G.; COREEA, V. Análise Espacial de Área. In DRUCK, S.; CARVALHO, M. S.; CÂMARA, G; MONTEIRO, A. M. V. **Análise Espacial de Dados Geográficos**. Brasília: EMBRAPA, 2004b. p.136-180.

CARVALHO, A. G. M. L.; LIMAYLLA, D. C. VILCHES, T. N. de ALMEIDA, G. B. MADALOSSO, G. de Assis, D. B. FORTALEZA, C. M. C. B. Spatial and sociodemographic factors associated with surgical site infection rates in hospitals in inner São Paulo State, Brazil. **Journal of Hospital Infection**, v. 108. 2021. pp. 181-184. DOI: 10.1016/j.jhin.2021.05.005.

CARAZATTO, P. Z. A. **Dinâmica de emergência e disseminação de enterobactérias resistentes a carbapenêmicos (CRE) e *Acinetobacter* baumannii multidroga-resistente no Brasil e no Estado de São Paulo: revisão sistemática e estudo de bases secundárias governamentais**. Tese de Doutorado. Programa de Pós-Graduação em Doenças Tropicais da Faculdade de Medicina de Botucatu, Universidade Estadual Paulista (UNESP). 2019. 166 p.

CASSINI, A., HÖGBERG, L. D., PLACHOURAS, D., QUATTROCCHI, A., HAXHA, A. SIMONSEN, G. S., COLOMB-COTINAT, M., KRETZSCHMAR, M. E., DEVLEESSCHAUWER, B., CECCHINI, M., OUAKRIM, D. A., OLIVEIRA, T. C., STRUELENS, M. J., SUETENS, C., MONNET, D. L. **Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European economic area in 2015: a population-level modelling analysis**. Lancet Infect Dis. v. 19, 2019. pp. 56–66.

Centers for Disease Control and Prevention. **Antibiotic Resistance Threats in the United States, 2019**. U.S. Department of Health and Human Services. 2019. Disponível em: <https://www.cdc.gov/antimicrobial-resistance/media/pdfs/2019-ar-threats-report-508.pdf>. Acesso em: 05 de agosto de 2024.

_____. **Antibiotic resistance threats in the United States**. Centers for Disease Control and Prevention, Atlanta, GA, 2013.
Centers for Disease Control and Prevention. **Available online**: Disponível em: <https://www.cdc.gov/healthcare-associated-infections/?CDC_AAref_Val=https://www.cdc.gov/hai/infectiontypes.html>. Acesso em: 30 de novembro de 2024.

CEVS – Centro Estadual de Vigilância em saúde RS. **Controle da disseminação de *Acinetobacter spp* resistentes a carbapenêmicos no município de Porto Alegre**. Disponível em: < https://www.cevs.rs.gov.br/upload/arquivos/201706/30122749-20120521095513manual-de-controle-da-disseminacao-do-acinectobacter.pdf>. Acesso em: 10 de outubro de 2022.

CHAN, H.-P.; SAMALA, R. K.; HADJIISKI, L. M., ZHOU, C.. Deep Learning in Medical Image Analysis. **PubMed Central**, 2020. doi: 10.1007/978-3-030-33128-3_1

CHANG, K.-T. **Introduction to Geographic Information Systems**. 9. ed. Boston: McGraw-Hill, 2019. 444p.

CHARNIAK, E. **An Introduction to Deep Learning**. The MIT Press. London, England, 2018. 174 p.

CHEN, T., GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System.** Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785-794. doi:10.1145/2939672.2939785.

CHIAVEGATTO FILHO, A. P.; BATISTA, A. F. M.; SANTOS, H. G. Data Leakage in Health Outcomes Prediction With *Machine learning*. Comment on “Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and *Machine learning*”. **JMIR Publications**. v. 23, n° 2, 2021.

CHNG, K. R.; LI, C.; BERTRAND, D.; NG, A. H. Q.; KWAH, J. S.; LOW, H. M.; TONG, C.; NATRAJAN, M.; ZHANG, M. H. XU, L.; KO, K. K. K.; HO, E. X. P.; AV-SHALOM, T. V.; TEO, J. W. P.; KHOR, C. C.; MetaSUB Consortium; CHEN, S. L.; MASON, C. E.; NG, O. T.; MARIMUTHU, K.; ANG, B.; NAGARAJAN, N. Cartography of opportunistic pathogens and antibiotic resistance genes in a tertiary hospital environment. **Nature Medicine**, v. 26, n. 6, p. 941-951, jun. 2020. DOI: https://doi.org/10.1038/s41591-020-0894-4.

CHOLLET, F. *Deep Learning com Python*. Rio de Janeiro: Alta Books, 2018.

CICCOLINI, M., DONKER, T., KÖCK, R., MIELKE, M., HENDRIX, R., JURKE, A., RAHAMAT-LANGENDOEN J., BECKER K., NIESTERS H. G., GRUNDMANN H., FRIEDRICH A.W. Infection prevention in a connected world: the case for a regional approach. **Int J Med Microbiol**. v. 303, 2013, pp. 380-387.

CICCOLINI, M.; DONKER, T.; GRUNDMANN, H.; BONTEN, M. J.; WOOLHOUSE, M. E. Efficient surveillance for healthcare-associated infections spreading between hospitals. **Proc Natl Acad Sci U S A**. v. 111, nº 6, 2014, pp. 2271-2276.

CLARK, P. J; EVANS, F. C. Distance to nearest neighbor as a measure of spatial relationships in populations. **Ecology**. v. 35. Issue 4, 1954. pp. 444-453.

CLARK, T. Can out-of-sample forecast comparisons help prevent overfitting? **Journal of Forecasting**. V. 23, 2 ed., 2004, pp. 115–139.

CNN BRASIL. **Veja como a conectividade aprimora processos hospitalares**. Publicado 11 de abril de 2024. Disponível em: <https://www.cnnbrasil.com.br/branded-content/nacional/veja-como-a-conectividade-aprimora-processos-hospitalares/ >. Acesso em: 10 de fevereiro de 2025.

COLLIGNON, P., BEGGS, J. J., WALSH, T. R., GANDRA, S. LAXMINARAYAN, R. Anthropological and socioeconomic factors contributing to global antimicrobial resistance: a univariate and multivariable analysis. **The Lancet Planetary Health**, v. 2, n. 9, pp. 398-405, 2018. Disponível em:

<https://www.thelancet.com/action/showPdf?pii=S2542-5196%2818%2930186-4>. Acesso em: 10 de novembro de 2024.

CONOVER, W. J. ***Practical Nonparametric Statistics***. 3. ed. John Wiley & Sons. New York, 1999.

DANDOLINI, B. W., BATISTA, L., de SOUZA, L. H. F., GALATO, D., & PIOVEZAN, A. P. **Uso Racional de Antibióticos: uma experiência para educação em saúde com escolares.** Ciência & Saúde Coletiva. v.17, n. 5, p. 1323-1331, 2012.

DEGENHARDT, F.; SEIFERT, S.; SZYMCZAK, S. Evaluation of variable selection methods for random forests and omics data sets. **Briefings in bioinformatics**, 20(2), 2019. pp. 492-503.

DELMELLE, E. Point Pattern Analysis. In: **International Encyclopedia of Human Geography**. Elsevier. 2009. pp.204-211.

DIÁRIO OFICIAL DO ESTADO DE SÃO PAULO. DECRETO Nº 51.433, de 28 de dezembro de 2006. Cria unidade na Coordenadoria de Regiões de Saúde, da Secretaria da Saúde, altera a denominação e dispõe sobre a reorganização das Direções Regionais de Saúde e dá providências correlatas. Disponível em: <http://dobuscadireta.imprensaoficial.com.br/default.aspx?DataPublicacao=20061229&Caderno=DOE-I&NumeroPagina=1>. Acesso em: 12 de outubro de 2022.

DONG, X., YU, Z., CAO, W., SHI, Y., MA, Q. **A survey on ensemble learning.** Frontiers Comput. Sci., vol. 14, no. 2, 2020. pp. 241-258.
doi: 1460 10.1007/s11704-019-8208-z.

DONKER, T., WALLINGA, J., SLACK, R., GRUNDMANN, H. Hospital networks and the dispersal of hospital-acquired pathogens by patient transfer. **PLoS One**. 4 ed. v. 7, 2012, pp. 1-8.

DRUCK, S.; CÂMARA, G.; MONTEIRO, A. M.; CARVALHO, M. S. **Análise Espacial de Dados Geográficos.** Brasília, EMBRAPA, 2004.

EGAN, J. P. Signal detection theory and ROC analysis, Series in Cognition and Perception. **Academic Press**, New York. 1975.
ESTEVA, A.; KUPREL, B.; NOVOA, R. A.; KO, J.; SWETTER, S. M.; BLAU, H. M.; THRUN, S. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 2017. pp. 115-118. Disponível em: <https://www.nature.com/articles/nature21056>. Acesso em: 25 de novembro de 2024.

FERETZAKIS, G., SAKAGIANNI, A. LOUPELIS, E., KALLES, D. MARTSOUKOU, M., SKARMOUTSOU, N., CHRISTOPOULOS, C., LADA, M., VELENTZA, A., PETROPOULOU, S., MICHELIDOU, S., KALDIS, V., CHATZIKYRIAKOU, R., DALAINAS, I., DIMITRELLOS, E. Using *Machine learning* to Predict Antimicrobial Resistance of *Acinetobacter* Baumannii, Klebsiella Pneumoniae and Pseudomonas Aeruginosa Strains. *Public Health and Informatics.* 2021. pp. 43-47
DOI: 10.3233/SHTI210117

FERNANDES, F. T.; CHIAVEGATTO FILHO, A. D. P. Predição de absenteísmo docente na rede pública com *machine learning*. **Revista de Saúde Pública**, v. 55, n. 23, 2021. Disponível em: <https://www.scielo.br/j/rsp/a/v6CY4PhVrH8wJwmSN396GFN/?lang=pt&format=pdf>. Acesso em: 10 de novembro de 2024.

FERNANDES, F. T. ***Machine learning* em saúde e segurança do trabalhador: perspectivas, desafios e aplicações.** Tese apresentada ao programa de pós-graduação em Saúde Pública da Faculdade de Saúde Pública da Universidade de São Paulo. 2021. 117p. Disponível em: <https://doi.org/10.11606/T.6.2021.tde-27012022-140548 >. Acesso em: 07 de Janeiro de 2025.

FISCHER, M. M.; WANG, J. **Spatial Data Analysis**: Models, Methods, and Techniques. Springer, New York, 2011. 82p.

FLOREK, P., ZAGDAŃSKI, A. Benchmarking state-of-the-art gradient boosting algorithms for classification. **Wrocław University of Science and Technology**, 2023. Disponível em: <https://arxiv.org/abs/2305.17094>. Acesso em: 16 de março de 2025.

FORSBERG, K. J.; REYES, A.; WANG, B.; SELLECK, E. M.; SOMMER, M. O.; DANTAS, G. The shared antibiotic resisto-me of soil bacteria and human pathogens. **Science**. 2012. pp.1107-1111.

FOTHERINGHAM, A. S., BRUNSDON, C., CHARLTON, M. **Quantitative Geography: Perspectives on Spatial Data Analysis**. Sage Publications, 2000. 288 p.

FREIRE, Maristela Pinheiro; RINALDI, Matteo; TERRABUIO, Debora Raquel Benedita; FURTADO, Mariane; PASQUINI, Zeno; BARTOLETTI, Michele; OLIVEIRA, Tiago Almeida de; NUNES, Nathalia Neves; LEMOS, Gabriela Takeshigue; MACCARO, Angelo; SINISCALCHI, Antonio; LAICI, Cristiana; CESCON, Matteo; DT'ALBUQUERQUE, Luiz Augusto Carneiro; MORELLI, Maria Cristina; SONG, Alice T. W.; ABDALA, Edson; VIALE, Pierluigi; CHIAVEGATTO FILHO, Alexandre Dias Porto; GIANNELLA, Maddalena. Prediction models for carbapenem-resistant Enterobacterales carriage at liver transplantation: a multicenter retrospective study. *Transplant Infectious Disease*, [S.l.], v. 24, n. 6, e13920, 2022. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/tid.13920>. Acesso em: 08 de abril de 2025.

FRIEDMAN, J. H. **Greedy Function Approximation: A Gradient Boosting Machine.** The Annals of Statistics, 29(5), 2001. pp. 1189-1232.

GARCIA, P. G., SILVA, I. A. R., DAMIANSE, L. A., OLIVEIRA, L. R. G., AZEVEDO, R. A. S. (2017). Prevalência de enterobactérias produtoras de *Klebsiella pneumoniae* carbapenemase em culturas de vigilância epidemiológica em unidade de terapia intensiva de um hospital de ensino de Minas Gerais. **HU Revista**. v. 43, pp. 199-203, 2017. Disponível em: <https://doi.org/10.34019/1982-8047.2017.v43.2744>. Acesso em: 10 de agosto de 2024.

GEISSER, S. Predictive Inference: An Introduction. London: **Chapman and Hall/CRC**, 1993.

GELBAND, H.; MILLER-PETRIE, M.; PANT, S.; GANDRA, S.; LEVINSON, J.; BARTER, D.; WHITE, A.; LAXMINARAYAN, R. The state of the Word's antibiotics. **One Health**. Disponível em: <https://onehealthtrust.org/publications/state_worlds_antibiotics_2015>. Acesso em: 10 de janeiro de 2025.

GÉRON, A. Hands-on *machine learning* with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. **O'Reilly Media, Inc.** 2017.

GETIS, A. Spatial Association, Measures of. *International Encyclopedia of the Social & Behavioral Sciences*, 2nd ed., Elsevier, 2015. pp. 100-105. Disponível em: <http://dx.doi.org/10.1016/B978-0-08-097086-8.72055-1>. Acesso em 10 de janeiro de 2025.

GIACOBBE, D. R.; MORA, S.; GIACOMINI, M.; BASSETTI, M. *Machine learning* and Multidrug-Resistant Gram-Negative Bacteria: An Interesting Combination for Current and Future Research. **Antibiotics**, V. 9, 54; 2020. Disponível em: <https://doi.org/10.3390/antibiotics9020054>. Acesso em: 10 de janeiro de 2025.

GIANFRANCESCO, M. A.; TAMANG, S.; YAZDANY, J. SCHMAJUK, G. Potential biases in *machine learning* algorithms using electronic health record data. **The Lancet Digital Health**, v. 1, n. 6, 2019. pp. 223–232.

GLOBAL HEALTH. Unpacking the relationship between antimicrobial resistance & climate change. 2024. Disponível em: <https://rabinmartin.com/insights/unpacking-the-relationship-between-antimicrobial-resistance-climate-change/>. Acesso em: 10 de março de 2025.

GOLDSTEN, B. A.; NAVAR, A. M.; CARTER, R. E.. Moving beyound regression techniques in cardiovascular risk prediction: Applying *machine learning* to address analytic challenges. Eur Heart J. 2017. pp. 1805-1814

GOLDSTEIN, A.; KAPELNER, A.; BLEICH, J. PITKIN, E.. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. **Journal of Computational and Graphical Statistics**, v. 24, n. 1, p. 44-65, 2015. DOI: 10.1080/10618600.2014.907095

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning.** MIT Press. London, 2016. 767p.

GOTELLI, J. N.; ELLISON, A. M. **Princípios de estatística em ecologia**. Tradução: Landeiro V. L. Porto Alegre, RS: Artmed, 2011. 528 p.

GOVERNO DO ESTADO DE SÃO PAULO. **Plano de Prevenção e Controle de Bactérias Multirresistentes (BMR) para os Hospitais do Estado de São Paulo**. 2016. Disponível em: <https://www.saude.sp.gov.br/resources/cve-centro-de-

vigilancia-epidemiologica/areas-de-vigilancia/infeccao-hospitalar/bmr/doc/ih16_bmr_plano.pdf>. Acesso em: 05 de janeiro de 2022.

_____. **Redes Regionais de Atenção à Saúde – RRAS**. Disponível em: <https://saude.sp.gov.br>. Acesso em: 26 de junho de 2024.

_____. **Regionais de Saúde.** Disponível em: <https://saude3.saude.sp.gov.br/departamentos-regionais-de-saude/regionais-de-saude/ >. Acesso em 02 de julho de 2024.

GREKOUSIS, G. ***Spatial Analysis Methods and Practice***. United Kingdom: Cambridge University Press, 2020. 518p.

GRUNDMANN, H. Towards a global antibiotic resistance surveillance system: a primer for a roadmap. Ups. **J. Med. Sci**. 119(2). 2014. pp. 87-97.

HAQUE, M.; SARTELLI, M.; MCKIMM, J.; ABU BAKAR, M. **Healthcare-associated infections – An overview. Infect. Drug Resist**. 2018, 11, 2321–2333. [CrossRef] [PubMed]

HARRIS, M., FASOLINO, T., IVANKOVIC, D., DAVIS, N. J., BROWNLEE, N. Genetic factors that contribute to antibiotic resistance through intrinsic and acquired bacterial genes in urinary tract infections. **Microorganisms**, v. 11, n. 6, p. 1407, 2023. Disponível em: <https://www.mdpi.com/2076-2607/11/6/1407>. Acesso em: 11 de abril de 2025.

HART, S. V; **Mapping Crime: Understanding Hot Spots.** CreateSpace Independent Publishing Platform. 2014, 78p.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. 2. ed. Springer. New York, 2008. 745 p.

HAUKE, J., KOSSOWSKI, T. (2011). Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. Quaestiones Geographicae, 30(2), 87-93. Disponível em: <https://sciendo.com/article/10.2478/v10117-011-0021-1>. Acesso em: 24 de julho de 2024.

HAWKINS, D. M. The problem of overfitting. **Journal of Chemical Information and Computer Sciences**. v. 44, 2004, pp. 1–12.

HAYKIN, S. **Neural networks and learning machines.** 3 ed. Pearson Education, New Jersey, 2009. 906 p.

HENDRIKSEN, R. S., MUNK, P., NJAGE, P., VAN BUNNIK, B, MCNALLY, L., LUKJANCENKO, O., RÖDER, T., NIEUWENHUIJSE, D., PEDERSEN, S. K., KJELDGAARD, J., KAAS, R. S., CLAUSEN, P. T. L. C., VOGT, J. K., LEEKITCHAROENPHON, P., VAN DE SCHANS, M. G. M., ZUIDEMA, T., HUSMAN, A. M. R., RASMUSSEN, S., PETERSEN, B., CONSORTIUM, The Global Sewage Surveillance project; AMID, C., COCHRANE, G., SICHERITZ-PONTEN, T., SCHMITT,

H., ALVAREZ, J. R. M. AIDARA-KANE, A., PAMP, S., J., LUND, O., HALD, T., WOOLHOUSE, M., KOOPMANS, M. P., VIGRE, H., PETERSEN, T. N., AARESTRUP, F. M. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. **Nature Communications**, v. 10, n. 1124, 2019. Disponível em: <https://www.nature.com/articles/s41467-019-08853-3>. Acesso em: 04 de janeiro de 2025.

HONG, T. P.; KUO, C. S.; CHI, S. C. Mining association rules from quantitative data. Intelligent data analysis, **IOS Press**, v. 3, n. 5, 1999. pp. 363–376.

HOSMER, D. W., LEMESHOW, S. **Applied Logistic Regression**. 2nd ed., New York, Wiley, 2000.

HUANG, K. T.; LEE, Y.W.; WANG, R. Y. Quality information and knowledge. New York: Prentice-Hall, 1999.

IZBICKI, R.; SANTOS, T. M. ***Machine learning* sob a ótica estatística: uma abordagem preditivista para a estatística com exemplos em R**. 2019. Disponível em: <http://www.rizbicki.ufscar.br/sml.pdf>. Acesso em: 18 de agosto de 2022.

IBANHES, L. C.; RIBEIRO, E. A. W.; GUIMARÃES, R. B.; PESSOTO, U. C.. Regionalização e as Redes Regionais de Atenção à Saúde no Estado de São Paulo. PEGADA - A Revista da Geografia do Trabalho, 2014 Disponível em:<https://revista.fct.unesp.br/index.php/pegada/article/view/2923>. Acesso em 10 de fevereiro de 2024.

IBGE. Instituto Brasileiro de Geografia e Estatística. Censo 2022 – População do Estado de São Paulo. 2022. Disponível em: <https://cidades.ibge.gov.br/brasil/sp/sao-paulo/panorama>. Acesso em: 25 de maio de 2024.

_____. **Malha municipal.** Disponível em: <https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais/15774-malhas.html>. Acesso em: 10 de agosto de 2021.

IZBICKI, R.; SANTOS, T. M. **Aprendizado de máquina: uma abordagem estatística.** Livro eletrônico. São Carlos, SP, 2020. 253 p.

JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data**. Prentice-Hall, Inc., 1988. 320p.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning with Applications in R.** Springer. New York, 2013. 426 p.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. 6 ed. Prentice Hall, 2007. 773p.

JORNAL DA USP. **Morrem no mundo 700 mil pessoas por ano vítimas de bactérias resistentes.** Disponível em: <https://jornal.usp.br/campus-ribeirao-preto/morrem-no-mundo-700-mil-pessoas-por-ano-vitimas-de-bacterias-resistentes/>. Acesso em:16 de maio de 2024.

KHALEDI, A.; WEIMANN, A.; SCHNIEDERJANS, M.; ASGARI, E.; KUO, T.-H.; OLIVER, A.; CABOT, G.; KOLA, A.; GASTMEIER, P.; HOGARDT, M.; JONAS, D.; MOFRAD, M. R. K.; BREMGES, A.; MCHARDY, A. C.; HÄUSSLER, S.. Predicting antimicrobial resistance in Pseudomonas aeruginosa with *machine learning*-enabled molecular diagnostics. **EMBO Molecular Medicine**, v. 12, n. 3, e10264, 2020. DOI: 10.15252/emmm.201910264.

KAUFMAN, S.; ROSSET, S.; PERLICH, C. Leakage in Data Mining: Formulation, Detection, and Avoidance. Presented at: Proceedings of the 17th **ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**; August 21-24, 2011, San Diego, CA, USA p. 556-563.

KAWAMOTO, M. T. **Análise de técnicas de distribuição espacial com padrões pontuais e aplicação a dados de acidentes de trânsito e a dados de dengue de Rio Claro–SP**. Dissertação (mestrado) – Universidade Estadual Paulista, Instituto de Biociências de Botucatu: Botucatu, SP, 2012.

KE, G., MENG, Q., FINLEY, T., WANG, T., CHEN, W., MA, W., YE, Q., LIU, T.-Y. **LightGBM: A Highly Efficient Gradient Boosting Decision Tree**. Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017. pp. 3149-3157.

KITABA, Abera A.; BONGER, Zelalem T.; BEYENE, Degefu; AYENEW, Zeleke; TSIGE, Estifanos; KEFALE, Tesfa A.; MEKONNEN, Yonas; TEKLU, Dejenie S.; SEYOUM, Elias; NEGERI, Abebe A. Antimicrobial resistance trends in clinical Escherichia coli and Klebsiella pneumoniae in Ethiopia. **African Journal of Laboratory Medicine**, v. 13, n. 1, p. 1–7, 2024. Disponível em: https://doi.org/10.4102/ajlm.v13i1.2268. Acesso em: 8 maio 2025.

KIFFER, C. R. V.; CUBA, G. T.; FORTALEZA, C. M. C. B.; PADOVEZE, M. C.; PIGNATARI, A. C. C. Exploratory model for estimating occupation-day costs associated to Hospital Related Infections based on data from national prevalence Project: IRAS Brasil Project. **Journal of Infection Control**. Ano IV. v. 4. n° 1. 2015.

KIMERLING, A. J.; BUCKLEY, A. R.; MUEHRCKE, P. C.; MUEHRCKE, J. O. **Map Use**: Reading, Analysis, Interpretation. Redlands: Esri Press Academic. 8 ed. 2016. 908p.

KUHN, M., JOHNSON, K. ***Applied Predictive Modeling***. Michigan, USA: Springer. 5 ed., 2016, 600 p.

LEE, J. *Statistics, Descriptive*. In: **International Encyclopedia of Human Geography.** Elsevier. pp.422-428. 2009. DOI:10.1016/b978-008044910-4.00534-4

LEGENZA, L.; MCNAIR, K.; GAO, S.; LACY, J. P.; OLSON, B. J.; FRITSCHE, T. R.; SCHULZ, L. T.; LAMURO, S.; SPRAY-LARSON, F.; SIDDIQUI, T.; ROSE, W. E. A geospatial approach to identify patterns of antibiotic susceptibility at a neighborhood level in Wisconsin. **Journal of Antibiotic Research**, 2023. Disponível em: <https://doi.org/10.1038/s41598-023-33895-5 >. Acesso em: 10 de janeiro de 2025.

LESAGE, J. P., PACE, R. K. **Introduction to Spatial Econometrics**. CRC Press, 1st Edition, 2009. 331p.
LEVINE, N. CrimeStat: A Spatial Statistical Program for the Analysis of Crime Incidents. **ACM SIGSPATIAL International**, Computer Science, Geography, 2017. DOI: 10.1007/978-3-319-17885-1_229

LIMA, C.C., BENJAMIM, S.C.C., SANTOS, R.F.S. **Bacterial resistance mechanismag a instdrugs: a review**. CuidArte, Enferm; v.11, n.1, p. 105-113, 2017. DOI: https://doi.org/10.34117/bjdv6n10-518

LOPATKIN, A. J., BENING, S. C., MANSON, A. L., STOKES, J. M., KOHANSKI, M. A., BADRAN, A. H., EARL, A. M., CHENEY, N. J., YANG, J. H., COLLINS, J. Clinically relevant mutations in core metabolic genes confer antibiotic resistance. **PubMed**, v. 371, 2021. DOI: https://doi.org/10.1126/science.aba0862.

LU, Y. Spatial Clustering, Detection and Analysis of. In: **International Encyclopedia of Human Geography**. Elsevier. 2009. pp.327-324.

LUNDBERG, S. M.; LEE, S. I. A Unified Approach to Interpreting Model Predictions. **Advances in Neural Information Processing Systems**, 2017. pp. 4768-4777. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>. Acesso em: 21 de dezembro de 2024.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, v. 5, 1943. pp. 115-133.

MACFADDEN, D. R., MCGOUGH, S. F., FISMAN, D., SANTILLANA, M., BROWNSTEIN, J. S. Antibiotic resistance increases with local temperature. **Nature Climate Change**. v. 8, n. 6, pp. 510–514, 2018. Disponível em: <https://pmc.ncbi.nlm.nih.gov/articles/PMC6201249/pdf/nihms-986473.pdf>. Acesso em: 08 de março de 2025.

MACQUEEN, B. "Some Methods for Classification and Analysis of Multivariate Observations." Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, 1, Berkeley, CA: University of California Press (1967), pp. 281-297.

MARINO, A., MANIACI, A., LENTINI, M., RONSIVALLE, S., NUNNARI, G., COCUZZA, S., PARISI, F. M., CACOPARDO, B., LAVALLE, S., LA VIA, L. The global burden of multidrug-resistant bacteria. **Epidemiologia**, Basel, v. 6, n. 2, p. 21, 2025. Disponível em: <https://doi.org/10.3390/epidemiologia6020021>. Acesso em: 09 maio 2025.

MARTÍNEZ-AGÜEROA, S.; MARQUESA, A. G.; MORA-JIMÉNEZA, I.; ALVÁREZ-RODRÍGUEZ, J. SOGUERO-RUIZA, C. Multimodal Interpretable Data-Driven Models for Early Prediction of Antimicrobial Multidrug Resistance Using Multivariate Time-Series. arXiv, 2024. Disponível em: <https://arxiv.org/abs/2402.06295>. Acesso em: 06 de janeiro de 2025.

MARTINS, C. R., FERNANDES, E. P., SILVA, D. H. Antibiotic Use and Resistance Patterns in São Paulo: A Geospatial Perspective. **Infectious Diseases Journal**, 2018.
MEURER, W. J.; TOLLES, J. Logistic regression diagnostics: understanding how well a model predicts outcomes. **JAMA Network**. 2017. pp.1068-9.

MIENYE, D., SUN, Y. **A survey of ensemble learning: concepts, algorithms, applications, and prospects**. IEEE Access, 2022. DOI:10.1109/ACCESS.2022.3207287

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Editora UFMG, Belo Horizonte, 2005, 297p.

MINISTÉRIO DA SAÚDE. **Regionalização.** Disponível em: <https://www.gov.br/saude/pt-br>. Acesso em: 26 de junho de 2024.

_____. **Padronização da nomenclatura do censo hospitalar.** Secretaria de Assistência à Saúde, Departamento de Sistemas e Redes Assistenciais. – 2.ed. revista – Brasília: Ministério da Saúde, 2002. Disponível em: <https://bvsms.saude.gov.br/bvs/publicacoes/padronizacao_censo.pdf>. Acesso em: 28 de junho de 2024.

MITCHELL, A. **The Esri Guide to GIS Analysis: Geographic Patterns and Relationships.** v. 1., 2 ed. ESRI Press, 2020.

MITCHELL, A.; GRIFFIN, L. S. ***The ESRI Guide to GIS Analysis: Spatial Measurements and Statistics***. v.2., 2 ed. ESRI Press, 2021.

MORAN, P. A. P. Notes on continuous stochastic phenomena. **Biometrika**. v. 37, n. ½, 1950. pp. 17-23

MORI, V., GRANT, G., HATTINGH, L. Evaluation of antimicrobial resistance surveillance data sources in primary care setting: a scoping review. **Family Practice**, v. 42, 2025. Disponível em: <https://doi.org/10.1093/fampra/cmaf013>. Acesso em: 7 de maio de 2025.

MOUBARECK, C. A.; HALAT, D. H. Insights into *Acinetobacter* baumannii: A Review of Microbiological, Virulence, and Resistance Traits in a Threatening Nosocomial Pathogen. **PubMed**. 2020. pp. 1-29.

MUNK, P., BRINCH, C., MOLLER, F. D., PETERSEN, T. N., HENDRIKSEN, R. S., SEYFARTH, A. M., KJELDGAARD, J. S., SVENDSEN, C. A., VAN BUNNIK, B., BERGLUND, F., GLOBAL SEWAGE SURVEILLANCE CONSORTIUM, LARSSON, D. G. J., KOOPMANS, M. WOOLHOUSE, M., AARESTRUP. F. M. Genomic analysis of sewage from 101 countries reveals global landscape of antimicrobial resistance. **Nature Communications**, v. 13, n. 1, p. 7251, 2022. DOI: https://doi.org/10.1038/s41467-022-34312-7.

NEKKAB, N.; ASTAGNEAU, P.; TEMIME, L.; CRÉPEY, P. Spread of hospital-acquired infections: A comparison of healthcare networks. **PubMed**. 2017.

NETER, J.; KUTNER, M. H.; NACHTSHEIM, C. J.; WASSERMAN. W. Applied linear statistical models. 4 ed. **WCB McGraw-Hill**, New York. 1996.

NGUYEN, M., LONG, S. W., MCDERMOTT, P. F., OLSEN, R. J., OLSON, R., STEVENS, R. L., TYSON, G. H., ZHAO, S., DAVIS, J. J. **Using *Machine learning* to Predict Antimicrobial MICs and Associated Genomic Features for Nontyphoidal Salmonella**. Journal of Clinical Microbiology. 2019. **DOI:** https://doi.org/10.1128/jcm.01260-18

NIELSEN, D. **Tree Boosting With XGBoost. Why Does XGBoost Win "Every" *Machine learning* Competition?** Master of Science in Physics and Mathematics. NTNU – Norwegian University of Science and Technology. 2016. 98p. Disponível em: <https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2433761/16128_FULLTEXT.pdf>. Acesso em: 02 de agosto de 2024.

NORDMANN, P.; CUZON, G.; NAAS, T. The real threat of KPC carbapenemase-producing bacteria. **Lancet Infect Dis**. 2009, vol. 9, pp. 321-331.

OLIVEIRA, L. Q., NEGREIROS, R. V., GOMES, C. B. S., SILVEIRA, H. L. SOUSA, A. O. B. Prevalência de bactérias multirresistentes em Unidade de Terapia Intensiva. **Revista Interdisciplinar em Saúde**. Cajazeiras, v. 7, pp. 2168-2181, 2020. DOI: 10.35621/23587490.v7.n1.p2168-2181

O'NEILL, J. Tackling Drug-Resistant Infections Globally: Final Report and Recommendations. Government of the United Kingdom. 2016

O'SULLIVAN, D.; UNWIN, D. J. **Geographic information analysis**. 2 ed. New Jersey: John Wiley Sons. 2010. 405p.

OLSON, R. S.; LA CAVA, W.; MUSTAHSAN, Z.; VARIK, A. MOORE, J. H. Data-driven Advice for Applying *Machine learning* to Bioinformatics Problems. **Biocomputing 2018**. 2018. pp. 192-203.

PATERSON, D.L. Resistance in gram-negative bacteria: Enterobacteriaceae. **Am J Infect Control**. 2006; v.34 (5 SUPPL.): pp. 20-28. https://doi.org/10.1016/j.ajic.2006.05.238

PEARCE, J. Regression, Linear and Nonlinear. In: KITCHIN, R.; THRIFT, N. (Ed.). *International Encyclopedia of Human Geography*. Oxford: Elsevier, 2009. p. 302–308. DOI: https://doi.org/10.1016/B978-008044910-4.00507-1.

PFEIFFER, D. U.; ROBINSON, T. P.; STEVENSON, M.; STEVENS, K. B.; ROGERS, D. J.; CLEMENTS, A. C. A. **Spatial Analysis in Epidemiology**. Oxford: Oxford University Press. 2008. 171p.

PILLONETTO, M.; JORDÃO, R. T. S.; ANDRAUS, G. S.; BÉRGAMO, R.; ROCHA, F. B.; ONISHI, M. C.; DE ALMEIDA, B. M. M.; NOGUEIRA, K. D. S.; DAL LIN, A.; DIAS, V. M. C. H; DE ABREU, A. L. The Experience of Implementing a National Antimicrobial Resistance Surveillance System in Brazil. **Public Health**. 2020.

PITTET, D.; ALLEGRANZI, B.; STORR, J.; BAGHERI, N. S.; DZIEKAN, G.; LEOTSAKOS, A.; DONALDSON, L. Infection control as a major World Health Organization priority for developing countries. **The Journal of Hospital Infection**. 2008. pp. 285-292.

PORTAL HOSPITAIS BRASIL. Infecção Hospitalar: problema ainda afeta 14% dos pacientes internados no Brasil. Publicado 24 de setembro de 2019. Disponível em: < https://portalhospitaisbrasil.com.br/infeccao-hospitalar-problema-ainda-afeta-14-dos-pacientes-internados-no-brasil/>. Acesso em: 30 de novembro de 2022.

POUTANEN, S. M.; LOUIE, M.; SIMOR, A. E. Risk Factors, Clinical Features and Outcome of *Acinetobacter* Bacteremia in Adults. **European Journal of Clinical Microbiology & Infectious Diseases**. v. 16, 1997. pp. 737-740.

PREFEITURA DE SÃO PAULO. **Entidades Públicas da RMSP**. Disponível em: <https://gestaourbana.prefeitura.sp.gov.br/marco-regulatorio/pdui/entidades-publicas-da-rmsp/>. Acesso em: 2 de dezembro de 2024.

PROKHORENKOVA, L. O., GUSEV, G., VOROBEV, A., DOROGUSH, A. V., GULIN, A. **CatBoost: unbiased boosting with categorical features.** Advances in Neural Information Processing Systems, v. 31, 2018. pp. 6638-6648.

QUEIROZ, E. R. S.; MEDRONHO, R. A. Spatial analysis of the incidence of Dengue, Zika and Chikungunya and socioeconomic determinants in the city of Rio de Janeiro, Brazil. **Epidemiology and Infection**, Cambridge, v. 149, p. e188, 2021. Disponível em: <https://www.cambridge.org/core/journals/epidemiology-and-infection/article/spatial-analysis-of-the-incidence-of-dengue-zika-and-chikungunya-and-socioeconomic-determinants-in-the-city-of-rio-de-janeiro-brazil/203F0D2364AB3EE7EBE3C311E98E3A26>. Acesso em: 09 abril 2023.

RAJKOMAR, A.; DEAN, J.; KOHANE, I.. "*Machine learning* in medicine. **New England Journal of Medicine**, v. 380, n.14, 2019. pp. 1347-1358. DOI: 10.1056/NEJMra1814259

RAMOS, M. C. A., da CRUZ, L. P., KISHIMA, V. C., POLLARA, W. M., de LIRA, A. C. O., COUTTOLENC, B. F. Avaliação de desempenho de hospitais que prestam atendimento pelo sistema público de saúde, Brasil. **Rev Saúde Pública.** n° 49, 2015. Disponível em: <https://doi.org/10.1590/S0034-8910.2015049005748>. Acesso em: 14 de julho de 2021.

RASCHKA, S. **Model Evaluation, Model Selection, and Algorithm Selection in *Machine learning***. **arXiv**, 2020. 49 p.

RASCHKA, S.; MIRJALILI, V. Python *machine learning*. **Packt Publishing Ltd**, 2017.

RASCHKA, S; MIRJALILI, V. **Python *Machine learning*: *Machine learning* and Deep Learning with Python, scikit-learn, and TensorFlow 2**. Packt Publishing. 3th Ed. 2019, 741 p.

RAWSON, T. M. MING, D., AHMAD, R., MOORE, L. S. P., HOLMES, A. H. Artificial intelligence can improve decision-making in infection management. **Nature Human Behaviour**, London, v. 3, pp. 543–545, 2019. DOI: https://doi.org/10.1038/s41562-019-0583-9.

REVELAS, A. **Healthcare-associated infections: A public health problem**. Niger. Med. J. 2012, 53, 59–64. [CrossRef] [PubMed]

RIBEIRO, E. A. W., GUIMARÃES, R. B., PESSOTO, U. C. **Regionalização e as Redes Regionais de Atenção à Saúde no Estado de São Paulo | Regionalização e as Redes Regionais de Atenção à Saúde no Estado de São Paulo.** PEGADA – A Revista da Geografia do Trabalho. Disponível em: <https://revista.fct.unesp.br/index.php/ pegada/article/view/2923>. Acesso em: 27 de junho de 2024.

RIPLEY, B. D. Spatial Statistics. **John Wiley & Sons**, New Jersey, 1981. 252 p.

ROCHA, D. J. P. G. **Abordagem Integrada para Predição de Resistência Antimicrobiana em *Corynebacterium sp*. Multirresistente**. Dissertação de Mestrado, Universidade Federal da Bahia, 2021. Disponível em: <https://repositorio.ufba.br/handle/ri/37478>. Acesso em: 09 de janeiro de 2025.

RODRIGUES, W.C., 2023. Teste de Kolmogorov-Smirnov. DivEs - Diversidade de Espécies v.4.21 (AntSoft Systems On Demand) - Guia do Usuário. Disponível em: <https://dives.ebras.bio.br>. Acesso em: 19 de julho de 2024.

RODRIGUES, M. M. S., BARRETO, B. D.; VINHAES, C. L.; ARAÚJO, M, P.; FUKUTANI E. R.; BERGAMASCHI, K. B.; KRISTKI, A.; CORDEIRO, M. S.; ROLLA, V. C.; STERLING, T. R.; QUEIROZ, A. T. L.; ANDRADE, B. B.. *Machine learning* algorithms using national registry data to predict loss to follow-up during tuberculosis treatment. **BMC Public Health**. 2024 May 23;24(1):1385. doi: 10.1186/s12889-024-18815-0. PMID: 38783264; PMCID: PMC11112756.

ROSA, S. C. L.; CARVALHO, N. I.; DUANI, H. Aprendizado de máquinas para predição de resistência microbiana. **Journal of Health Informatics**, 2024. Disponível em: <https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/1264/591>. Acesso em: 10 de janeiro de 2025.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain, **Psychological Review**. 1958. pp. 386–408.

ROSVALL, M.; BERGSTROM, C. T. Maps of random walks on complex networks reveal community structure. **Proceedings of the National Academy of Sciences of the United States of America**. 2008.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. (1986-10-09). Learning representations by back-propagating errors. **Nature**. 1986.

SAMARASINGHE, S. Neural networks for applied sciences and engineering: from fundamentals to complex pattern recognition. **Auerbach Publications**. 2006.

SAKAGIANNI, A., KOUFOPOULOU, C., FERETZAKIS, G., KALLES, D., VERYKIOS, V. S., MYRIANTHEFS, P., FILDISIS, G. Using Machine learning to Predict Antimicrobial Resistance – A Literature Review. **PubMed**. v. 12, n. 3, 2023. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/36978319/>. Acesso em: 10 de janeiro de 2025.

SANTOS, H. G. Comparação da performance de algoritmos de *machine learning* para a análise preditiva em saúde pública e medicina. Tese de Doutorado. Programa de Pós-Graduação em Epidemiologia da Faculdade de Saúde Pública da Universidade de São Paulo. 2018. 206 p.

SANTOS, H. G.; NASCIMENTO, C. F.; IZBICKI, R.; DUARTE, Y. A. O.; CHIAVEGATTO FILHO, A. D. P. *Machine learning* para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil. **Cad Saúde Pública**. 2019

SCHMIDHEINY, K. The bootstrap. Short Guides to Microeconometrics. Unversität Basel. Disponível em: <https://www.schmidheiny.name/teaching/bootstrap.pdf>. Acesso em: 01 de setembro de 2022.

SHWARTZ-ZIV, R.; ARMON, A. *Tabular Data: Deep Learning is Not All You Need*. Information Fusion, v. 81, p. 84–90, 2022. DOI: https://doi.org/10.1016/j.inffus.2021.11.011.

SEADE. Fundação Sistema Estadual de Análise de Dados. Disponível em: <https://www.seade.gov.br/>. Acesso em: 24 de fevereiro de 2022.

Secretaria de Estado da Saúde de São Paulo. **Centro de Vigilância Sanitária. Gestão da Visa: Mapas de Saúde e Vigilância Sanitária**. Disponível em: < http://www.cvs.saude.sp.gov.br/prog_det.asp?te_codigo=36&pr_codigo=25>. Acesso em: 14 de julho de 2022.

_____. **Departamentos Regionais de Saúde.** Disponível em: <https://www.saude.sp.gov.br/ses/institucional/departamentos-regionais-de-saude/?page=1>. Acesso em: 05 de outubro de 2022.

SIEGEL, S., CASTELLAN, N. J., Jr.. **Nonparametric statistics for the behavioral sciences** (2nd ed.). Mcgraw-Hill Book Company, 1988. Disponível em: <https://journals.sagepub.com/doi/10.1177/014662168901300212>. Acesso em: 10 de outubro de 2024.

DA SILVA, J. A., PEREIRA, M. J., ALMEIDA, M. P. (2020). *Machine learning* applications in public health: A review of current approaches and future directions. Journal of Biomedical Informatics, 104, 103394. doi:10.1016/j.jbi.2020.103394.

SAHARMAN, Y. R., KARUNIAWATI, A., SEVERIN, J. A., VERBRUGH, H. A. Infections and antimicrobial resistance in intensive care units in lower-middle income countries: a scoping review. **Antimicrobial Resistance & Infection Control**, v. 10, n. 1, p. 22, 2021. Dispónível em: <https://link.springer.com/article/10.1186/s13756-020-00871-x>. Acesso em: 20 de janeiro de 2025.

SHIMAKURA, S. **Estatística II**. Departamento de Estatística da Universidade Federal do Paraná. 2006. Disponível em: <http://leg.ufpr.br/~silvia/CE003/node74.html>. Acesso em: 30 de agosto de 2024.

SILVEIRA, Z. P.; MALINKIEWICZ, A.; MENEZES, M. B.; SOUSA, E. O.; FREITAS, L. M. A.; CAZEIRO, C. C.; SILVA, D. R. C.; CARNEIRO, E. N. A.; FARIAS, D. C. S.; CRUZ, L. P. S.; ORTA, B. H. S.; MACEDO, V. C. A automedicação com antibióticos e as repercussões na resistência bacteriana. **Revista Ibero-Americana de Humanidades, Ciências e Educação**, São Paulo, v. 9, n. 07, p. 545-556, jul. 2023. ISSN 2675-3375. Disponível em: <https://doi.org/10.51891/rease.v9i7.10653>. Acesso em: 22 de novembro de 2024.

SILVERMAN, B. W. **Density estimation for statistics and data analysis**. London: Chapman and Hall, 1986.

SINGER, R. S., WARD, M. P., MALDONADO, G. Can landscape ecology untangle the complexity of antibiotic resistance? **Nature Publishing Group**. 2006. pp. 943-952. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/17109031/>. Acesso em: 22 de dezembro de 2024.

SINGH, N.; BHATNAGAR, S. *Machine learning* for Prediction of Drug Targets in Microbe Associated Cardiovascular Diseases by Incorporating Host-pathogen Interaction Network Parameters. **Molecular Informatics**. 2021. Disponível em: <https://doi.org/10.1002/minf.202100115>. Acesso em: 26 de dezembro de 2024.

SORRE, M. Principes de cartographie applique a l'écologie humaine. **Social Science & Medicine**, v.12, D, p. 238-50, 1978.

SPETS, P., EBERT, K., DINNÈTZ, P. Spatial analysis of antimicrobial resistance in the environment. A systematic review Authors. Geospatial Health 2023, v. 18:1168. https://doi.org/10.4081/gh.2023.1168

STOESSER, N., SHEPPARD, A. E., SHAKYA, M., THORSON, S., GIESS, A., KELLY, D., PETO, T. E., CROOK, D. W., WALKER, A. S. (2013). Dynamics of antimicrobial resistance in Escherichia coli during long-term asymptomatic carriage in humans. Clinical Infectious Diseases, v. 58, 2013. pp. 799-806.

TAO, Z., LIU, H., LI, S., DING, Z., FU, Y. **Marginalized multiview ensemble clustering**, IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 2, pp. 600_611, Feb. 2020, doi: 10.1109/TNNLS.2019.2906867

TACCONELLI, E., CARRARA, E., SAVOLDI, A., HARBARTH, S., MENDELSON, M., MONNET, D.L., PULCINI, C. KAHLMETER, G. KLUYTMANS, J. CARMELI, Y. OUELLETTE, M. OUTTERSON, K. PATEL, J., CAVALERI, M., COX, E. M., HOUCHENS, C. R., GRAYSON, M. L., HANSEN, P., SINGH, N., THEURETZBACHER, U., MAGRINI, N., WHO Pathogens Priority List Working Group. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis**. PubMed.** 2018. pp. 318–327.

TANG, R., LUO, R., TANG, S., SONG, H. CHEN, X. Machine learning in predicting antimicrobial resistance: a systematic review and meta-analysis. **International Journal of Antimicrobial Agents**, v. 60, p. 106684, 2022. DOI: 10.1016/j.ijantimicag.2022.106684.

TEIXEIRA, B. C., TOPORCOV, T. N., CHIARAVALLOTI-NETO, F., CHIAVEGATTO FILHO, A. D. P. (2023). Spatial Clusters of Cancer Mortality in Brazil: A Machine Learning Modeling Approach. **International Journal of Public Health**, v. 68, 1604789. Dispónível em: <https://doi.org/10.3389/ijph.2023.1604789>. Acesso em: 10 de agosto de 2024.

TOBLER, W. R. A computer movie simulating urban growth in the Detroit region. **Economic Geography**. Taylor & Francis, Ltd. v. 46, 1970. pp. 234-240.

TUFTE, Edward R. **The visual display of quantitative information.** 2. ed. Cheshire, Connecticut: Graphics Press, 2007. 197 p.

VAPNIK, V.; GOLOWICH, S.; SMOLASUPPORT, A. Support vector method for function approximation, regression estimation and signal processing. v. 4. 1996.

VILCHES, T. N.; BONESSO, M. F.; GUERRA, H. M. FORTALEZA, C. M. C. B. PARK, A. W. FERREIRA, C.P. The role of intra and inter-hospital patient transfer in the dissemination of heathcare-associated multidrug-resistant pathogens. **Epidemics**, v. 26, 2019, pp. 104–115.

VERVIER, K.; MAHÉ, P.; TOURNOUD, M.; VEYRIERAS, J.-B.; VERT, J.-P.. Large-scale *machine learning* for metagenomics sequence classification. **PubMed** – Bioinformatics, v. 32, n. 7, p. 1023-1032, 2016. DOI: 10.1093/bioinformatics/btv683.

VOLPATO, Gilson; BARRETO, Rodrigo. Estatística sem dor. 2º ed. Botucatu: Best Writing, 2016.

WALLER, L. A.; GOTWAY, C. A. **Applied Spatial Statistics for Public Health Data**. New Jersey: John Wiley & Sons. 2004. 494p.

WANG, F. Factor Analysis and Principal-Components Analysis. **Elsevier** – International Encyclopedia of Human. 2009. pp. 1-7.

WANG, Y. **Wetlands and habitatis**. CRC Press. 2 ed. v. III. 2020. 294p.

WANG. C.; VENKATESH, S. **Optimal Stopping and Effective Machine Complexity in Learning**. Advances in NIPS, 1984. pp. 303–310.

WANG, T. HANSEN, K. R., LOVING, J., PASCHALIDIS, I. C., VAN AGGELEN, H., SIMHON, E.. Predicting Antimicrobial Resistance in the Intensive Care Unit. **arXiv preprint** arXiv:2111.03575, 2021. Disponível em: <https://arxiv.org/abs/2111.03575>. Acesso em: 18 de dezembro de 2024.

WANG, X., PATIL, N., LI, F., WANG, Z., ZHAN, H., SCHMIDT, D., THOMPSON, P., GUO, Y., LANDERSDORFER, C. B., SHEN, H., PELEG, A. Y., LI, J., SONG, J.,

PmxPred: A data-driven approach for the identification of active polymyxin analogues against gram-negative bacteria. **Computers in Biology and Medicine**, v. 168, 107681, 2024. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0010482523011460>. Acesso em: 11 de abril de 2025.

WARTENBERG, D. Multivariate spatial correlation: a method for exploratory geographical analysis. **Geographical Analysis**. n.17. 1985, pp. 263–283.

WICHMANN, R. M.; FERNANDES, F. T.; CHIAVEGATTO FILHO, A. D. P.; CICONELLE, A. C. M.; DE BRITO, A. M. E. S.; NUNES, B. P.; SILVA, D. L. E.; ANSCHAU, F.; RODRIGUES, H. C.; ROCHA, H. A. L.; DOS REIS, J. C. B.; CAVALCANTE, L. O.; DE OLIVEIRA, L. P.; ANDRADE, L. S. S. ; NASI, L. A.; FELIX, M. M.; MIMICA, M. J.; ARAUJO, M. E. A.; ARNONI, M. V. ; VIANNA, R. B.. Improving the performance of *machine learning* algorithms for health outcomes predictions in multicentric cohorts. **Scientific Reports**, v. 13, pp. 1022, 2023.

WHO – World Health Organization. **WHO Guidelines for Hand Hygiene in Health Care (Advanced Draft)**. Geneva: World Health Organization, 2006.

_____. **Guidelines for the prevention and control of carbapenem-resistant Enterobacteriaceae, Acinetobacter baumannii and Pseudomonas aeruginosa in health care facilities**. World Health Organization, Geneva, Switzerland, 2017.

_____. **WHO publishes list of bacteria for which new antibiotics are urgently needed**. 2017. Disponível em: <https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed>. Acesso em: 20 de maio de 2024.

_____. **Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics**. Geneva: WHO, 2017. Disponível em: <https://www.who.int/publications/i/item/WHO-EMP-IAU-2017.12>. Acesso em: 20 de maio de 2024.

_____. **New report calls for urgent action to avert antimicrobial resistance crisis [homepage on the Internet]. Jt News Release**. 2019, p. 1-4. Disponível em: <https://www.who.int/news/item/29-04-2019-new-report-calls-for-urgent-action-to-avert-antimicrobialresistance-crisis>. Acesso em: 10 de fevereiro de 2025.

_____. **Global Antimicrobial Resistance and Use Surveillance System (GLASS) Report**. 2021. Disponível em: <https://apps.who.int/iris/bitstream/handle/10665/341666/9789240027336-eng.pdf>. Acesso em: 19 de setembro de 2022.

_____. **World Hand Hygiene Day: Key facts and figures on antimicrobial resistance and healthcare-associated infections**. Geneva, 2023. Disponível em: <https://www.who.int/campaigns/world-hand-hygiene-day/key-facts-and-figures>. Acesso em: 26 abril 2025.

WONG, D. W. S.; LEE, J.. **Statistical Analysis with ArcView GIS and ArcGIS**. Wiley & Sons, Inc., New York, 2005. 464 p.

WU, Y.-L.; HU, X.-Q.; WU, D.-Q.; LI, R.-J.; WANG, X.-P.; ZHANG, J.; LIU, Z.; CHU, W.-W.; ZHU, X.; ZHANG, W.-H.; ZHAO, X.; GUAN, Z.-S.; JIANG, Y.-L.; WU, J.-F.; CUI, Z.; ZHANG, J.; LI, J.; WANG, R.-M.; SHEN, S.-H.; CAI, C.-Y.; ZHU, H.-B.; JIANG, Q.; ZHANG, J.; NIU, J.-L.; XIONG, X.-P.; TIAN, Z.; ZHANG, J.-S.; ZHANG, J.-L.; TANG, L.-L.; LIU, A.-Y.; WANG, C.-X.; NI, M.-Z.; JIANG, J.-J.; YANG, X.-Y.; YANG, M.; ZHOU, Q. Prevalence and risk factors for colonisation and infection with carbapenem-resistant Enterobacterales in intensive care units: A prospective multicentre study. **PubMed**, v. 87, n. 4, p. 299-308, 2023. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/37480701/>. Acesso em: 12 de dezembro de 2024.

YI, J., KIM, K. H. **Identification and infection control of carbapenem-resistant Enterobacterales in intensive care units.** Acute Crit Care. 2021. pp.175184.

YOO, H., SOKHANSANJ, B., BROWN, J. R., ROSEN, G.. Predicting Anti-microbial Resistance using Large Language Models. *arXiv preprint arXiv:2401.00642*, 2024. Disponível em: <https://arxiv.org/abs/2401.00642>. Acesso em: 13 de janeiro de 2025.

ZHANG, H., WANG, Y., & LI, Y. (2022). Prediction of methicillin-resistant *Staphylococcus aureus* infections in hospitals using neural networks. *BMC Infectious Diseases*, 22(1), 211.

ZHANG, J., MUCS, D., NORINDER, U., SVENSSON, F. **LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity – Application to the Tox21 and Mutagenicity Datasets**. Journal of Chemical Information and Modeling v. 59 Ed. 10, 2019. pp. 4150-4158.

ZAR, J. H. (1972). Significance testing of the Spearman rank correlation coefficient. Journal of the American Statistical Association, 67(339), pp. 578-580.
Disponível em: < https://www.jstor.org/stable/2284441>. Acesso em: 24 de julho de 2024.

ZIMLICHMAN, E.; HENDERSON, D.; TAMIR, O.; FRANZ, C.; SONG, P.; YAMIN, C. K.; KEOHANE, C.; DENHAM, C. R.; BATES, D. B. Health care-associated infections a meta-analysis of costs and financial impact on the US health care system. JAMA Intern. Med., 173, 2039-2046, 2013.

ZORZETTO, R. **O avanço das superbactérias**. Revista PESQUISA FAPESP v. 335, 2019. pp. 13-17. Disponível em: <https://revistapesquisa.fapesp.br/wp-content/uploads/2024/01/012-017_capa-infeccoes_335-Parte-1.pdf>. Acesso em: 23 de dezembro de 2023.

ZWEIG, M. H.; CAMPBELL, G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. **Clin Chem**. 1993. pp.561-577.

ANEXO A – RRAS, DRS E REGIÃO DE SAÚDE

RRAS	DRS	Região de Saúde	Número de Municípios
01	Grande São Paulo	Grande ABC	7
02	Grande São Paulo	Alto do Tietê	11
03	Grande São Paulo	Franco da Rocha	5
04	Grande São Paulo	Mananciais	8
05	Grande São Paulo	Rota dos Bandeirantes	7
06	Grande São Paulo	São Paulo	1
07	Baixada Santista Registro	Baixada Santista	9
		Vale do Ribeira	15
08	Sorocaba	Itapetininga	13
		Itapeva	15
		Sorocaba	20
09	Bauru	Vale do Jurumirim	17
		Bauru	18
		Polo Cuesta	13
		Jau	12
		Lins	8
10	Marília	Adamantina	10
		Assis	13
		Marília	19
		Ourinhos	12
		Tupã	8
11	Presidente Prudente	Alta Paulista	12
		Alta Sorocabana	19
		Alto Capivari	5
		Extremo Oeste Paulista	5
		Pontal do Paranapanema	4
12	Araçatuba São José do Rio Preto	Central do DRS II	11
		Dos Lagos do DRS II	12
		Dos Consórcios do DRS II	17
		Catanduva	19
		Santa Fé do Sul	6
		Jales	16
		Fernandópolis	13
		São José do Rio Preto	20
		José Bonifácio	11
		Votuporanga	17
13	Araraquara Barretos Franca Ribeirão Preto	Central do DRS III	8
		Centro Oeste do DRS III	5
		Norte do DRS III	5
		Coração do DRS III	6
		Norte-Barretos	10
		Sul-Barretos	8
		Três Colinas	10
		Alta Anhanguera	6
		Alta Mogiana	6
		Horizonte Verde	9
		Aquífero Guarani	10
		Vale das Cachoeiras	7

14	Piracicaba	Araras	5
		Limeira	4
		Piracicaba	11
		Rio Claro	6
15	Campinas São João da Boa Vista	Campinas	11
		Oeste VII	11
		Baixa Mogiana	4
		Mantiqueira	8
		Rio Pardo	8
16	Campinas	Bragança	11
		Jundiaí	9
17	Taubaté	Alto Vale do Paraíba	8
		Circuito Fé – Vale Histórico	17
		Litoral Norte	4
		Vale do Paraíba – Região Serrana	10
TOTAL			**645**

Fonte: Secretaria de Estado da Saúde de São Paulo (2022).

ANEXO B – CÓDIGO DE *MACHINE LEARNING* – CLASSIFICAÇÃO

```
#1. Instalação de pacotes adicionais
# ver descrição dos pacotes na seção de importação
!pip install dfply
!pip install scikit-plot
!pip install graphviz
!pip install dtreeviz

#2. Importação de bibliotecas e pacotes
import pandas as pd # para processamento de bancos de dados
import numpy as np # para processamento numérico de bancos de dados
import matplotlib # para geração de gráficos
import matplotlib.pyplot as plt # configurações adicionais para os gráficos a serem gerados

# informamos ao Python que estamos usando um notebook e que os gráficos devem ser exibidos
nele
%matplotlib inline

import seaborn as sns #alternativa para a matplotlib para geração de gráficos

# SCIKIT-LEARN
from sklearn.metrics import roc_curve, auc
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.model_selection import KFold, cross_val_score
from sklearn.model_selection import train_test_split, GridSearchCV, RandomizedSearchCV
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report
from sklearn.calibration import CalibratedClassifierCV, calibration_curve
from sklearn.metrics import brier_score_loss, precision_score, recall_score, f1_score, roc_auc_score
from sklearn import preprocessing
from sklearn.preprocessing import StandardScaler

from dfply import *  # para importar os comandos da biblioteca dfply

import scikitplot as skplt

from dtreeviz.trees import *

import warnings # ignorando os warnings emitidos pelo Python
warnings.filterwarnings("ignore")

# configurações adicionais de figuras
plt.rcParams["figure.figsize"] = [10, 5]
plt.style.use("fivethirtyeight")
%config InlineBackend.figure_format = 'retina'
from matplotlib import rc
rc('font',**{'family':'sans-serif','sans-serif':['DejaVu Sans'],'size':10})
rc('mathtext',**{'default':'regular'})

np.random.seed(42)  # semente de aleatoriedade

#3. Obtendo o conjunto de dados
dataset =
pd.read_csv('C:/Users/Renata/Documents/pythonrenata/planilha_mun_2011_2019_classificacao.csv',
sep = ';')

#3.1 Verificando o conjunto de dados
```

```
dataset.shape
dataset.head(5)
dataset.info()

# Verificar se há valores ausentes (NaN) ou valores infinitos
import numpy as np
import pandas as pd

# Carregar os dados
X_test =
pd.read_csv('C:/Users/Renata/Documents/pythonrenata/planilha_mun_2011_2019_classificacao.csv',
sep=';')

# Verificar valores NaN nos dados de teste
nan_check = X_test.isna().any()

# Verificar valores infinitos apenas nas colunas numéricas
numeric_cols = X_test.select_dtypes(include=[np.number])
inf_check = np.isinf(numeric_cols).any()

# Exibir os resultados
print("Valores NaN:")
print(nan_check)

print("\nValores infinitos:")
print(inf_check)

print(dataset.columns)

#4. Preparação do conjunto de dados (pré-processamento)
#4.1 Filtrando o conjunto de dados
cvd_data = (dataset >>
            drop('GEOCODE7d')
           )
cvd_data.head(5)
cvd_data.shape

#4.2 Para as variáveis categóricas iremos criar dummies
cvd_data.head(10).T
cvd_data[['EnteR']] = cvd_data[['EnteR']].apply(preprocessing.LabelEncoder().fit_transform)
cvd_data.head(10).T

#4.3 Obtendo os conjuntos de treino e teste
variaveis_preditoras = cvd_data.iloc[:, cvd_data.columns != 'AcinR'] # separamos as nossas variáveis
preditoras do nosso desfecho/target ==> conjunto X
classe = cvd_data.iloc[:, cvd_data.columns == 'AcinR']

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(variaveis_preditoras,
                                        classe,
                                        stratify = classe,   #ver
                                        train_size = 0.70,
                                        random_state = 42)

variaveis_preditoras.head(10).T
classe.head(20).T
++X_train.columns
#Variáveis selecionadas
X_train = X_train.loc[:,['Leitos','LetUti','PacDia', 'Nhosp', #'EnteT','EnteS','TxCRE', 'EnteR',
                    'Pop','Ano','UTIs','EnergR','EnergC','PIB','IDH2010', #,'Riqueza'
```

```
                    'DistKm','DDem']]

X_test = X_test.loc[:,['Leitos','LetUti','PacDia', 'Nhosp', #'EnteT','EnteS','TxCRE', 'EnteR',
                    'Pop','Ano','UTIs','EnergR','EnergC','PIB','IDH2010', #,'Riqueza'
                    'DistKm','DDem']]

X_train_columns = X_train.columns
X_test_columns = X_test.columns

#4.4 As variáveis contínuas serão padronizadas
rom sklearn.compose import ColumnTransformer
#Variáveis selecionadas
continuous_cols = ['Leitos','LetUti','PacDia', 'Nhosp',
                    'Pop','Ano','UTIs','EnergR','EnergC','PIB','IDH2010', #'Riqueza',
                    'DistKm','DDem']


def setScaler():
  ct = ColumnTransformer([
        ('scaler', StandardScaler(), continuous_cols)
    ], remainder='passthrough'
  )
  return ct

scaler = setScaler()

scaler.fit(X_train)

X_train = scaler.transform(X_train)
X_test = scaler.transform(X_test)

pd.set_option('display.float_format', lambda x: '%.2f' % x)
# a padronização retorna os dados em formato array.
X_train = pd.DataFrame(X_train, columns=X_train_columns)
X_test = pd.DataFrame(X_test, columns=X_test_columns)
X_train.head()
X_test.describe()
# transformando a variável target: Y --> 1 e N --> 0
le = preprocessing.LabelEncoder()
le.fit(y_train)
y_train = le.transform(y_train)
y_test = le.transform(y_test)
y_train

#5. Função auxiliar RunModel
def runModel(model, X_train, y_train, X_test, y_test, confusion_matrix=True, normalizeCM=False,
roc=True, plot_calibration=True, random_state=42, title="", pos_label=1):
clf = model
    name = title
    clf.fit(X_train, y_train)
    y_pred = clf.predict(X_test)

    if hasattr(clf, "predict_proba"):
        prob_pos = clf.predict_proba(X_test)
    else:  # usar decision function
        prob_pos = clf.decision_function(X_test)
        prob_pos = (prob_pos - prob_pos.min()) / (prob_pos.max() - prob_pos.min())
    if confusion_matrix:
       skplt.metrics.plot_confusion_matrix(y_test, y_pred, normalize=normalizeCM, title=name)
    if roc:
```

```
      skplt.metrics.plot_roc(y_test, prob_pos, plot_micro=False, plot_macro=False, classes_to_plot=[1],
title=name,figsize=(10,10))


    prob_pos = prob_pos[:,1]
    clf_score = brier_score_loss(y_test, prob_pos, pos_label=pos_label)
    print("%s:" % name)
    print("\tBrier: %1.3f" % (clf_score))
    print("\tROC(AUC) %1.3f" % roc_auc_score(y_test, prob_pos))
    print("\tPrecision: %1.3f" % precision_score(y_test, y_pred))
    print("\tRecall: %1.3f" % recall_score(y_test, y_pred))
    print("\tF1: %1.3f\n" % f1_score(y_test, y_pred))

    if plot_calibration:

      fraction_of_positives, mean_predicted_value = \
              calibration_curve(y_test, prob_pos, n_bins=10)
      plt.rcParams.update({'font.size': 22})
      plt.rc('legend',**{'fontsize':22})
      fig = plt.figure(3, figsize=(10, 10))
      ax1 = plt.subplot2grid((3, 1), (0, 0), rowspan=2)
      ax2 = plt.subplot2grid((3, 1), (2, 0))
      ax1.plot([0, 1], [0, 1], "k:", label="Perfeitamente calibrado",)
      ax1.plot(mean_predicted_value, fraction_of_positives, "s-",
                  label="%s (%1.3f)" % (name, clf_score))

      ax2.hist(prob_pos, range=(0, 1), bins=10, label=name,
                  histtype="step", lw=2)

      ax1.set_ylabel("Fração de positivos")
      ax1.set_ylim([-0.05, 1.05])
      ax1.legend(loc="lower right")
      ax1.set_title('Gráfico de Calibração  (reliability curve)')

      ax2.set_xlabel("Valor médio predito")
      ax2.set_ylabel("Quantidade")
      ax2.legend(loc="upper center", ncol=2)

      for item in ([ax1.title, ax1.xaxis.label, ax1.yaxis.label] +
            ax1.get_xticklabels() + ax1.get_yticklabels()):
        item.set_fontsize(22)

      for item in ([ax2.title, ax2.xaxis.label, ax2.yaxis.label] +
            ax2.get_xticklabels() + ax2.get_yticklabels()):
        item.set_fontsize(22)

      plt.tight_layout()
      plt.show()

# Função de best scores
def report(results, n_top=3):
    for i in range(1, n_top + 1):
        candidates = np.flatnonzero(results['rank_test_score'] == i)
        for candidate in candidates:
            print("Model with rank: {0}".format(i))
            print("Mean validation score: {0:.3f} (std: {1:.3f})".format(
                  results['mean_test_score'][candidate],
                  results['std_test_score'][candidate]))
            print("Parameters: {0}".format(results['params'][candidate]))
            print("")
```

```
help(runModel)

#6. Execução dos algoritmos de machine learning

#6.1 Random Forest
# modelo random forest
rf = RandomForestClassifier(random_state=42, verbose=1)

np.random.seed(42)

# Número de árvores no Random Forest
n_estimators = [int(x) for x in np.linspace(start = 100, stop = 1000, num = 5)]
# Número de features a serem consideradas a cada split
max_features = ['log2', 'sqrt']
# Número máximo de níveis na árvore
max_depth = [int(x) for x in np.linspace(5, 20, num = 5)]
# Número mínimo de amostras necessárias para dividir um nó
min_samples_split = [2, 5, 10]
# Número mínimo de amostras necessárias em cada leaf node
min_samples_leaf = [2, 4]
# Método de seleção das amostras para treinamento de cada árvore
bootstrap = [True, False]
# Criação do param grid
param_grid = {'n_estimators': n_estimators,
                'max_features': max_features,
                'max_depth': max_depth#,
                }
cv_rf = RandomizedSearchCV(rf, n_iter=50, cv=3, verbose=1, param_distributions=param_grid,n_jobs
= -1)

# otimizando os hiperparâmetros
cv_rf.fit(X_train, y_train)
# melhores estimadores
rf = cv_rf.best_estimator_
rf
rf = RandomForestClassifier(max_depth=8, max_features='log2', n_estimators=550,
            random_state=42, verbose=1)
# performance do modelo
runModel(rf, X_train, y_train, X_test, y_test, title="Random Forest")

import shap
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# Treinando o modelo Random Forest otimizado
rf.fit(X_train, y_train)

# Criando o interpretador SHAP
explainer = shap.TreeExplainer(rf)
shap_values = explainer.shap_values(X_test)

# Selecionar os SHAP values da classe positiva (1), em caso de classificação binária
shap_class_values = shap_values[1]

# Calcular a importância média absoluta por variável
importance = np.abs(shap_class_values).mean(axis=0)

# Converter para porcentagem
importance_percent = 100 * importance / importance.sum()
```

```
# Ordenar as variáveis pela importância
sorted_indices = np.argsort(importance_percent)[::-1]
sorted_feature_names = X_test.columns[sorted_indices]
sorted_importance_percent = importance_percent[sorted_indices]

# Selecionar top 10
top_10_features = sorted_feature_names[:10]
top_10_importance = sorted_importance_percent[:10]

# Plotar gráfico personalizado com porcentagem
plt.figure(figsize=(10, 6))
bars = plt.barh(top_10_features[::-1], top_10_importance[::-1], color='blue')
plt.xlabel('Importância (%)', fontsize=18)
#plt.ylabel('Variáveis', fontsize=14)
#plt.title('Importância das Variáveis - Random Forest (SHAP)', fontsize=16)
plt.xticks(fontsize=18)
plt.yticks(fontsize=18)
plt.grid(True, axis='x', linestyle='--', alpha=0.7)
plt.xlim(0, 35)

# Adicionar valores percentuais ao lado das barras
#for bar, value in zip(bars, top_10_importance[::-1]):
#    plt.text(value + 1, bar.get_y() + bar.get_height() / 2, f'{value:.1f}%', va='center', fontsize=12)

plt.tight_layout()
plt.show()

#6.2 XGBOOST
!pip install xgboost
import xgboost as xgb
model = xgb.XGBClassifier()
runModel(model, X_train, y_train, X_test, y_test, title="XGBoost")
# Importações
!pip install shap
import shap
import numpy as np
import matplotlib.pyplot as plt

# Ajustar o modelo aos dados
model.fit(X_train, y_train)

# Criar o interpretador SHAP
explainer = shap.Explainer(model, X_train)
shap_values = explainer(X_test)

# Calcular a importância média absoluta
importance = np.abs(shap_values.values).mean(axis=0)

# Converter em porcentagem
importance_percent = 100 * importance / importance.sum()

# Ordenar variáveis por importância
sorted_indices = np.argsort(importance_percent)[::-1]
sorted_feature_names = X_train.columns[sorted_indices]
sorted_importance_percent = importance_percent[sorted_indices]

# Selecionar top 10
top_10_features = sorted_feature_names[:10]
top_10_importance = sorted_importance_percent[:10]
```

```
# Plotar gráfico manual com personalização
plt.figure(figsize=(10, 6))
bars = plt.barh(top_10_features[::-1], top_10_importance[::-1], color='blue')
plt.xlabel('Importância (%)', fontsize=18)
#plt.ylabel('Variáveis', fontsize=18)
#plt.title('Importância das Variáveis - XGBoost (SHAP)', fontsize=16)
plt.xticks(fontsize=18)
plt.yticks(fontsize=18)
plt.xlim(0, 35)
plt.grid(True, axis='x', linestyle='--', alpha=0.7)

# Adicionar os valores percentuais ao lado das barras
#for bar, value in zip(bars, top_10_importance[::-1]):
#    plt.text(value + 1, bar.get_y() + bar.get_height() / 2, f'{value:.1f}%', va='center', fontsize=12)

plt.tight_layout()
plt.show()

#6.3 LightGBM
!pip install lightgbm
import lightgbm as lgb
clf_lgbm = lgb.LGBMClassifier()
runModel(clf_lgbm, X_train, y_train, X_test, y_test, title="LightGBM")

# Importações necessárias
import shap
import numpy as np
import matplotlib.pyplot as plt

# Treinar o modelo
clf_lgbm.fit(X_train, y_train)

# Criar interpretador SHAP
explainer = shap.TreeExplainer(clf_lgbm)
shap_values = explainer.shap_values(X_test)

# Para classificação binária, selecionamos os shap_values da classe 1
shap_class_values = shap_values[1]

# Calcular a média absoluta dos valores SHAP por variável
importance = np.abs(shap_class_values).mean(axis=0)

# Converter para porcentagem
importance_percent = 100 * importance / importance.sum()

# Ordenar as variáveis pela importância
sorted_indices = np.argsort(importance_percent)[::-1]
sorted_feature_names = X_test.columns[sorted_indices]
sorted_importance_percent = importance_percent[sorted_indices]

# Selecionar as 10 variáveis mais importantes
top_10_features = sorted_feature_names[:10]
top_10_importance = sorted_importance_percent[:10]

# Plotar gráfico de barras com porcentagens
plt.figure(figsize=(10, 6))
bars = plt.barh(top_10_features[::-1], top_10_importance[::-1], color='blue')
plt.xlabel('Importância (%)', fontsize=18)
#plt.ylabel('Variáveis', fontsize=14)
```

```
#plt.title('Importância das Variáveis - LightGBM (SHAP)', fontsize=16)
plt.xticks(fontsize=18)
plt.yticks(fontsize=18)
plt.xlim(0, 35)
plt.grid(True, axis='x', linestyle='--', alpha=0.7)

# Adicionar os valores percentuais nas barras
#for bar, value in zip(bars, top_10_importance[::-1]):
#    plt.text(value + 1, bar.get_y() + bar.get_height() / 2, f'{value:.1f}%', va='center', fontsize=12)

plt.tight_layout()
plt.show()

#6.4 Catboost
!pip install catboost
from catboost import CatBoostClassifier
clf_catboost = CatBoostClassifier()
runModel(clf_catboost, X_train, y_train, X_test, y_test, title="Catboost")

import shap
import numpy as np
import matplotlib.pyplot as plt

# Treinando o modelo CatBoost
clf_catboost.fit(X_train, y_train)

# Criando o interpretador SHAP
explainer = shap.TreeExplainer(clf_catboost)
shap_values = explainer.shap_values(X_test)

# Calcular a importância média absoluta por variável
importance = np.abs(shap_values).mean(axis=0)

# Converter para porcentagem
importance_percent = 100 * importance / importance.sum()

# Ordenar variáveis por importância
sorted_indices = np.argsort(importance_percent)[::-1]
sorted_feature_names = X_test.columns[sorted_indices]
sorted_importance_percent = importance_percent[sorted_indices]

# Selecionar as 10 variáveis mais importantes
top_10_features = sorted_feature_names[:10]
top_10_importance = sorted_importance_percent[:10]

# Criar gráfico com fonte maior (18)
plt.figure(figsize=(12, 7))
bars = plt.barh(top_10_features[::-1], top_10_importance[::-1], color='blue')
plt.xlabel('Importância (%)', fontsize=18)
plt.xticks(fontsize=18)
plt.yticks(fontsize=18)
plt.xlim(0, 35)
plt.grid(True, axis='x', linestyle='--', alpha=0.7)

# Adicionando os valores percentuais ao lado das barras

plt.tight_layout()
plt.show()

#6.5 Redes Neurais
```

```
import numpy as np
from sklearn.neural_network import MLPClassifier
from sklearn.model_selection import RandomizedSearchCV
mlp = MLPClassifier(random_state=42)
np.random.seed(42)

# Número de camadas e neurônios em cada camada oculta
hidden_layer_sizes = [(x,) for x in range(100, 1000, 200)]

# Função de ativação
activation = ['logistic', 'tanh', 'relu']

# Número máximo de iterações
max_iter = [200, 400, 600]

# Taxa de aprendizado inicial
learning_rate_init = [0.001, 0.01, 0.1]

# Criação do param grid
param_grid = {'hidden_layer_sizes': hidden_layer_sizes,
              'activation': activation,
              'max_iter': max_iter,
              'learning_rate_init': learning_rate_init}

cv_mlp = RandomizedSearchCV(mlp, n_iter=50, cv=3, verbose=1, param_distributions=param_grid,
n_jobs=-1)

# Otimizando os hiperparâmetros
cv_mlp.fit(X_train, y_train)

# Melhores estimadores
mlp = cv_mlp.best_estimator_
mlp

# Modificações ENTERO
mlp = MLPClassifier(hidden_layer_sizes=(500,), activation='relu', max_iter=600,
learning_rate_init=0.001, random_state=42)

# Performance do modelo
runModel(mlp, X_train, y_train, X_test, y_test, title="Neural Network")

import shap
import numpy as np
import matplotlib.pyplot as plt

# Ajustando o modelo MLP otimizado
mlp.fit(X_train, y_train)

# Reduzindo o número de amostras de fundo
X_train_sampled = shap.sample(X_train, 100)

# Criando o interpretador SHAP
explainer = shap.KernelExplainer(mlp.predict_proba, X_train_sampled)

# Calculando os valores SHAP para o conjunto de teste
shap_values = explainer.shap_values(X_test)

# Seleciona SHAP values da classe 1 (em classificação binária)
shap_values_class = shap_values[1]
```

```
# Calcular a importância média absoluta por variável
importance = np.abs(shap_values_class).mean(axis=0)

# Converter para porcentagem
importance_percent = 100 * importance / importance.sum()

# Ordenar variáveis por importância
sorted_indices = np.argsort(importance_percent)[::-1]
sorted_feature_names = X_train.columns[sorted_indices]
sorted_importance_percent = importance_percent[sorted_indices]

# Selecionar as 10 variáveis mais importantes
top_10_features = sorted_feature_names[:10]
top_10_importance = sorted_importance_percent[:10]

# Criar gráfico com fonte 18
plt.figure(figsize=(12, 7))
bars = plt.barh(top_10_features[::-1], top_10_importance[::-1], color='blue')
plt.xlabel('Importância (%)', fontsize=18)
#plt.ylabel('Variáveis', fontsize=18)
#plt.title('Importância das Variáveis - Redes Neurais (SHAP)', fontsize=18)
plt.xticks(fontsize=18)
plt.yticks(fontsize=18)
plt.xlim(0, 35)
plt.grid(True, axis='x', linestyle='--', alpha=0.7)

# Adicionar valores percentuais nas barras
#for bar, value in zip(bars, top_10_importance[::-1]):
#    plt.text(value + 1, bar.get_y() + bar.get_height() / 2, f'{value:.1f}%',
#             va='center', fontsize=18)

plt.tight_layout()
plt.show()
```

ANEXO C – CÓDIGO DE *MACHINE LEARNING* – REGRESSÃO

```
#Instalação dos pacotes adicionais
!pip install dfply
!pip install yellowbrick
!pip install scikit-plot
!pip install graphviz
!pip install dtreeviz

import pandas as pd # para processamento de bancos de dados
import numpy as np # para processamento numérico de bancos de dados
from dfply import *  # para importar os comandos da biblioteca dfply
import matplotlib.pyplot as plt # para geração de gráficos
from matplotlib import rc  # configurações adicionais para os gráficos a serem gerados

# informamos ao Python que estamos usando um notebook e que os gráficos devem ser exibidos
nele
%matplotlib inline
import seaborn as sns #alternativa para a matplotlib para geração de gráficos

# definimos o estilo dos gráficos
# mais estilos em https://matplotlib.org/3.1.1/gallery/#style-sheets
plt.style.use("fivethirtyeight")
%config InlineBackend.figure_format = 'retina' # formato das imagens
rc('font',**{'family':'sans-serif','sans-serif':['DejaVu Sans'],'size':10}) #fonte utilizada
rc('mathtext',**{'default':'regular'})

import warnings   # ignorando os warnings emitidos pelo Python
warnings.filterwarnings("ignore")

import operator  # para ordenação do zip

np.random.seed(42)  # semente de aleatoriedade

from sklearn.model_selection import train_test_split  # importamos a funcionalidade de split do
conjunto de dados em treino/teste

from sklearn.metrics import mean_squared_error, r2_score  # métricas de performance para modelos
de regressão

#Importando o conjunto de dados como um Pandas DataFrame
banco = pd.read_csv('C:/Users/Renata/Documents/pythonrenata/planilha_mun_2011_2019.csv', sep
= ';')
banco.info()
banco.head()
pd.set_option('display.float_format', lambda x: '%.3f' % x)
banco.describe()

#Separar conjunto de dados em treinamento e teste
# variável de interesse
outcome = banco >> select(X.AcinR)
#Retirar todas as variáveis de acinetobacter e enterobactérias
banco.drop(['GEOCODE7d', 'AcinT', 'AcinS', 'AcinR',
            'EnteT','EnteS', 'EnteR', 'TxCRE', 'TxCrab', 'Riqueza'], axis = 1, inplace = True)

X_train, X_test, y_train, y_test = train_test_split(banco, outcome, test_size=0.3)
X_train.shape
X_test.shape

#Pré-processamento dos dados de treinamento
```

```
import scipy.stats
corr=X_train.corr(method='spearman')
plt.figure(figsize=(15, 15))
sns.heatmap(corr, vmax=.8, linewidths=0.01,
        square=True,annot=True,cmap='YlGnBu',linecolor="white")
plt.title('Correlation between features')
plt.show()

#Filtrar preditores com variância nula
# importamos a função VarianceThreshold
from sklearn.feature_selection import VarianceThreshold
# Escolha o limiar aceitável de variância
threshold = 0

selector = VarianceThreshold(threshold)
selector.fit_transform(X_train)
for i,s in enumerate(selector.get_support()):
  if s:
    print(X_train.columns[i] + " - manter " + "["+ str(selector.variances_[i]) + "]")
  else:
    print("*** " + X_train.columns[i] + " - remover " + "["+ str(selector.variances_[i]) + "]")

constant_features = [
   feat for feat in X_train.columns if X_train[feat].std() == 0
]
constant_features

#Padronizar os dados de treino e teste
# importamos a funcao StandardScaler para padronização dos dados
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
sc.fit(X_train)
sc.mean_
sc.var_

# padronizamos o conjunto de treinamento
X_train = pd.DataFrame(sc.transform(X_train), columns=X_train.columns)
# padronizamos o conjunto de teste
X_test = pd.DataFrame(sc.transform(X_test), columns=X_test.columns)
X_train.describe()
X_test.describe()

#Random Forest Regressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import RandomizedSearchCV, GridSearchCV
# modelo random forest
rf = RandomForestRegressor(random_state=42)
#Hiperparâmetros a serem otimizados

# Verificar se há valores ausentes (NaN) ou valores infinitos.

import numpy as np
import pandas as pd

# Verifique valores NaN nos dados de teste
nan_check = pd.DataFrame(X_test).isna().any()

# Verifique valores infinitos nos dados de teste
inf_check = np.isinf(X_test).any()
```

```
print("Valores NaN:")
print(nan_check)

# Número de árvores no Random Forest
n_estimators = [int(x) for x in np.linspace(start = 5, stop = 600, num = 3)] # modificações minhas
# Número de features a serem consideradas a cada split
max_features = ['auto', 'sqrt']
# Número máximo de níveis na árvore
max_depth = [int(x) for x in np.linspace(5, 30, num = 3)]
# Número mínimo de amostras necessárias para dividir um nó
min_samples_split = [2, 5, 10]
# Número mínimo de amostras necessárias em cada leaf node
min_samples_leaf = [1, 2, 4]
# Método de seleção das amostras para treinamento de cada árvore
bootstrap = [True]

# Criação do param grid
param_grid_rf = {'n_estimators': n_estimators,
            'max_features': max_features,
            'max_depth': max_depth,
            'min_samples_split': min_samples_split,
            'min_samples_leaf': min_samples_leaf,
            'bootstrap': bootstrap}

# otimizando os hiperparâmetros
rf_random = RandomizedSearchCV(estimator = rf,
                           param_distributions = param_grid_rf,
                           n_iter = 50, ### número de avaliações do Random Grid
                           random_state = 42,
                           cv = 3,
                           verbose = 2,
                           n_jobs = -1)

# executando a busca
rf_random = rf_random.fit(X_train, y_train)

# resultados da busca
print('Melhor score: %s' % rf_random.best_score_)
print('Melhores hiperparâmetros: %s' % rf_random.best_params_)

#selecionando os resultados da busca
rf = RandomForestRegressor(n_estimators = 600,
                           min_samples_split = 2,
                           min_samples_leaf = 1,
                           max_features = 'sqrt',
                           max_depth = 5,
                           bootstrap = True,
                           random_state=42)  # alterando o valor do random_state ele altera os valores

# treinando o modelo otimizado
rf.fit(X_train, y_train)

#Medidas de performance
pred_train_rf= rf.predict(X_train)
print('RMSE (treino):', np.sqrt(mean_squared_error(y_train,pred_train_rf)))
print('R² (treino):', r2_score(y_train, pred_train_rf))

pred_test_rf= rf.predict(X_test)
print('RMSE (teste):',np.sqrt(mean_squared_error(y_test,pred_test_rf)))
print('R² (teste):',r2_score(y_test, pred_test_rf))
```

```
pred_test_rf = rf.predict(X_test)

# Arredondando os valores para inteiros
pred_test_int = pred_test_rf.astype(int)

#print('Previsões para o conjunto de teste:', pred_test_rf)

# Imprimindo as previsões como números inteiros
print('Previsões para o conjunto de teste (números inteiros):', pred_test_int)

# gráfico de dispersão: valor real vs valor predito
plt.figure(figsize=(15,10))
plt.scatter(y_test, pred_test_rf, c='crimson')

p1 = max(max(pred_test_rf), max(np.array(y_test)[0]))
p2 = min(min(pred_test_rf), min(np.array(y_test)[0]))
plt.title("Real vs Predito")
plt.plot([p1, p2], [p1, p2])
plt.xlabel('True Values', fontsize=15)
plt.ylabel('Predictions', fontsize=15)
plt.axis('equal')
plt.show()

Listar os valores preditos
for valor in pred_test_rf:
    print(valor)

# importância de variáveis com SHAP
import shap
import pandas as pd
import matplotlib.pyplot as plt

# Explicador e cálculo dos valores SHAP
explainer = shap.Explainer(rf, X_test, check_additivity=False)
shap_values = explainer(X_test, check_additivity=False)

# Calcula a média dos valores absolutos dos SHAP por variável
shap_values_abs = np.abs(shap_values.values)
mean_shap_importance = shap_values_abs.mean(axis=0)

# Converte para porcentagem
importance_percent = 100 * mean_shap_importance / mean_shap_importance.sum()

# Cria DataFrame com nomes das variáveis e ordena
importance_df = pd.DataFrame({
    'Variable': X_test.columns,
    'Importance (%)': importance_percent
}).sort_values(by='Importance (%)', ascending=False)

# Seleciona as 10 mais importantes
top_10 = importance_df.head(10)

# Plota o gráfico
plt.figure(figsize=(10, 6))
plt.barh(top_10['Variable'][::-1], top_10['Importance (%)'][::-1], color='blue')
plt.xlabel('Importância (%)', fontsize=18)
#plt.title('Importância das Variáveis - Random Forest (SHAP)', fontsize=16)
plt.xticks(fontsize=20)
plt.yticks(fontsize=20)
plt.xlim(0, 30)
```

```
plt.tight_layout()
plt.show()

#Redes Neurais

import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.preprocessing import StandardScaler
import tensorflow as tf
from tensorflow import keras
from tensorflow.keras import layers

# Os dados de treino e teste já foram divididos em: X_train, X_test, y_train e y_test

# Pré-processamento dos dados
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Criar uma rede neural
model = keras.Sequential([
    layers.Input(shape=X_train_scaled.shape[1]),
    layers.Dense(64, activation='relu'),
    layers.Dense(32, activation='relu'),
    layers.Dense(1)  # Camada de saída para previsão
])

# Compilar o modelo
model.compile(optimizer='adam', loss='mean_squared_error')

# Treinar a rede neural
model.fit(X_train_scaled, y_train, epochs=100, batch_size=32, validation_split=0.2, verbose=2)

# Avaliar o modelo nos dados de treinamento
pred_train_nn = model.predict(X_train_scaled)
print('RMSE (treino):', np.sqrt(mean_squared_error(y_train, pred_train_nn)))
print('R² (treino):', r2_score(y_train, pred_train_nn))

# Avaliar o modelo nos dados de teste
X_test_scaled = scaler.transform(X_test)
pred_test_nn = model.predict(X_test_scaled)
print('RMSE (teste):', np.sqrt(mean_squared_error(y_test, pred_test_nn)))
print('R² (teste):', r2_score(y_test, pred_test_nn))

# Plotar resultados
plt.figure(figsize=(15, 10))
plt.scatter(y_test, pred_test_nn, c='crimson')

p1 = max(max(pred_test_nn), max(np.array(y_test)[0]))
p2 = min(min(pred_test_nn), min(np.array(y_test)[0]))
plt.title("Real vs Predito")
plt.plot([p1, p2], [p1, p2])
plt.xlabel('True Values', fontsize=15)
plt.ylabel('Predictions', fontsize=15)
plt.axis('equal')
plt.show()

mport shap
```

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

# Calcular os valores SHAP
explainer = shap.DeepExplainer(model, X_train_scaled)
shap_values = explainer.shap_values(X_test_scaled)

# Selecionar os SHAP values da classe desejada (por exemplo, classe 0)
shap_values_class = shap_values[0]

# Calcular a importância média absoluta por variável
importance = np.abs(shap_values_class).mean(axis=0)

# Calcular importância em porcentagem
importance_percent = 100 * importance / importance.sum()

# Obter nomes das variáveis ordenados
sorted_indices = np.argsort(importance_percent)[::-1]
sorted_feature_names = X_train.columns[sorted_indices]
sorted_importance_percent = importance_percent[sorted_indices]

# Selecionar as 10 variáveis mais importantes
top_10_features = sorted_feature_names[:10]
top_10_importance = sorted_importance_percent[:10]

# Criar gráfico manual em porcentagem
plt.figure(figsize=(10, 6))
plt.barh(top_10_features[::-1], top_10_importance[::-1], color='blue')
plt.xlabel('Importância (%)', fontsize=20)
#plt.title('Importância das Variáveis - Redes Neurais (SHAP)', fontsize=16)
plt.xticks(fontsize=20)
plt.yticks(fontsize=20)
plt.xlim(0, 30)
plt.tight_layout()
plt.show()

#XGBoost
from xgboost import XGBRegressor
from sklearn.model_selection import RandomizedSearchCV
import numpy as np

# modelo XGBoost
xgb = XGBRegressor(random_state=42)

# Definindo os hiperparâmetros
param_grid_xgb = {
    'n_estimators': [int(x) for x in np.linspace(start=5, stop=600, num=3)],
    'max_depth': [6],
    'learning_rate': [0.001, 0.01, 0.1],
    'subsample': [0.3],
    'colsample_bytree': [0.5, 0.75, 1.0],
    'reg_alpha': [0],
    'reg_lambda': [1],
    }
# Realizando a busca aleatória
xgb_random = RandomizedSearchCV(estimator=xgb,
                                param_distributions=param_grid_xgb,
                                n_iter=50,
                                random_state=42,
```

```
                    cv=3,
                    verbose=2,
                    n_jobs=-1)

# Executando a busca aleatória
xgb_random.fit(X_train, y_train)

# Resultados da busca
print('Melhor score: %s' % xgb_random.best_score_)
print('Melhores hiperparâmetros: %s' % xgb_random.best_params_)

# Selecionando apenas os melhores resultados
xgb = XGBRegressor(**xgb_random.best_params_, random_state=42)

# Treinando o modelo otimizado
xgb.fit(X_train, y_train)

# Avaliando o desempenho do modelo nos dados de treinamento
pred_train_xgb = xgb.predict(X_train)
print('RMSE (treino):', np.sqrt(mean_squared_error(y_train, pred_train_xgb)))
print('R² (treino):', r2_score(y_train, pred_train_xgb))

# Avaliando o desempenho do modelo nos dados de teste
pred_test_xgb = xgb.predict(X_test)
print('RMSE (teste):', np.sqrt(mean_squared_error(y_test, pred_test_xgb)))
print('R² (teste):', r2_score(y_test, pred_test_xgb)) # Gráfico de dispersão: valor real vs valor predito
plt.figure(figsize=(15,10))
plt.scatter(y_test, pred_test_xgb, c='crimson')

# Determinar os limites dos eixos
p1 = max(max(pred_test_xgb), max(np.array(y_test)[0]))
p2 = min(min(pred_test_xgb), min(np.array(y_test)[0]))

# Plotar a linha diagonal
plt.plot([p1, p2], [p1, p2])

# Adicionar título e rótulos dos eixos
plt.title("Real vs Predito")
plt.xlabel('True Values', fontsize=15)
plt.ylabel('Predictions', fontsize=15)

# Ajustar proporções dos eixos
plt.axis('equal')

# Exibir o gráfico
plt.show()

#Importância das variáveis
import shap
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from xgboost import XGBRegressor

# Treinar o modelo XGBoost com os melhores hiperparâmetros encontrados
xgb = XGBRegressor(**xgb_random.best_params_, random_state=42)
xgb.fit(X_train, y_train)

# Calcular os valores SHAP
explainer = shap.Explainer(xgb, X_train)
```

```
shap_values = explainer.shap_values(X_test)

# Calcular importância média absoluta
importance = np.abs(shap_values).mean(axis=0)

# Converter em porcentagem
importance_percent = 100 * importance / importance.sum()

# Ordenar
sorted_indices = np.argsort(importance_percent)[::-1]
sorted_feature_names = X_train.columns[sorted_indices]
sorted_importance_percent = importance_percent[sorted_indices]

# Top 10
top_10_features = sorted_feature_names[:10]
top_10_importance = sorted_importance_percent[:10]

# Aumentar fonte
plt.figure(figsize=(10, 6))
plt.barh(top_10_features[::-1], top_10_importance[::-1], color='blue')
plt.xlabel('Importância (%)', fontsize=20)
#plt.ylabel('Variáveis', fontsize=14)
#plt.title('Importância das Variáveis - XGBoost (SHAP)', fontsize=16)
plt.xticks(fontsize=20)
plt.yticks(fontsize=20)
plt.xlim(0, 30)
plt.tight_layout()
plt.show()

#LightGBM
from lightgbm import LGBMRegressor
from sklearn.model_selection import RandomizedSearchCV
import numpy as np
# Modelo LightGBM
lgbm = LGBMRegressor(random_state=42)

# Definindo os hiperparâmetros
param_grid_lgbm = {
                'n_estimators': [int(x) for x in np.linspace(start=5, stop=600, num=3)],
                'max_depth': [5, 10, 15],
                'learning_rate': [0.001, 0.01, 0.1],
                'subsample': [0.5, 0.75, 1.0],
                'colsample_bytree': [0.5, 0.75, 1.0],
                'reg_alpha': [0, 0.001, 0.01, 0.1, 1, 10],
                'reg_lambda': [0, 0.001, 0.01, 0.1, 1, 10],
               }
# Realizando a busca aleatória
lgbm_random = RandomizedSearchCV(estimator=lgbm,
                            param_distributions=param_grid_lgbm,
                            n_iter=50,
                            random_state=42,
                            cv=3,
                            verbose=2,
                            n_jobs=-1)

# Executando a busca aleatória
lgbm_random.fit(X_train, y_train)

# Resultados da busca
print('Melhor score: %s' % lgbm_random.best_score_)
```

```
print('Melhores hiperparâmetros: %s' % lgbm_random.best_params_)

# Selecionando apenas os melhores resultados
lgbm = LGBMRegressor(**lgbm_random.best_params_, random_state=42)

Treinando o modelo otimizado
lgbm.fit(X_train, y_train)

# Avaliando o desempenho do modelo nos dados de treinamento
pred_train_lgbm = lgbm.predict(X_train)
print('RMSE (treino):', np.sqrt(mean_squared_error(y_train, pred_train_lgbm)))
print('R² (treino):', r2_score(y_train, pred_train_lgbm))

# Avaliando o desempenho do modelo nos dados de teste
pred_test_lgbm = lgbm.predict(X_test)
print('RMSE (teste):', np.sqrt(mean_squared_error(y_test, pred_test_lgbm)))
print('R² (teste):', r2_score(y_test, pred_test_lgbm))

# Gráfico de dispersão: valor real vs valor predito
plt.figure(figsize=(15,10))
plt.scatter(y_test, pred_test_lgbm, c='crimson')

# Determinar os limites dos eixos
p1 = max(max(pred_test_lgbm), max(np.array(y_test)[0]))
p2 = min(min(pred_test_lgbm), min(np.array(y_test)[0]))

# Plotar a linha diagonal
plt.plot([p1, p2], [p1, p2])

# Adicionar título e rótulos dos eixos
plt.title("Real vs Predito")
plt.xlabel('True Values', fontsize=15)
plt.ylabel('Predictions', fontsize=15)

# Ajustar proporções dos eixos
plt.axis('equal')

# Exibir o gráfico
plt.show()

#Importância das variáveis
import shap
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from lightgbm import LGBMRegressor

# Treinar o modelo LightGBM com os melhores hiperparâmetros
lgbm = LGBMRegressor(**lgbm_random.best_params_, random_state=42)
lgbm.fit(X_train, y_train)

# Calcular os valores SHAP
explainer = shap.TreeExplainer(lgbm)
shap_values = explainer.shap_values(X_test)

# Calcular a média dos valores absolutos dos SHAP por variável
importance = np.abs(shap_values).mean(axis=0)

# Converter para porcentagem
importance_percent = 100 * importance / importance.sum()
```

```
# Ordenar as variáveis pela importância
sorted_indices = np.argsort(importance_percent)[::-1]
sorted_feature_names = X_train.columns[sorted_indices]
sorted_importance_percent = importance_percent[sorted_indices]

# Selecionar as 10 variáveis mais importantes
top_10_features = sorted_feature_names[:10]
top_10_importance = sorted_importance_percent[:10]

# Plotar gráfico com fonte maior
plt.figure(figsize=(10, 6))
plt.barh(top_10_features[::-1], top_10_importance[::-1], color='blue')
plt.xlabel('Importância (%)', fontsize=20)
#plt.ylabel('Variáveis', fontsize=14)
#plt.title('Importância das Variáveis - LightGBM (SHAP)', fontsize=16)
plt.xticks(fontsize=20)
plt.yticks(fontsize=20)
plt.xlim(0, 30)
plt.tight_layout()
plt.show()

#Catboost
from catboost import CatBoostRegressor
from sklearn.model_selection import RandomizedSearchCV
import numpy as np
# Modelo CatBoost
catboost = CatBoostRegressor(random_state=42, verbose=False)

# Definindo os hiperparâmetros
param_grid_catboost = {
                    'iterations': [100, 200, 300],
                    'learning_rate': [0.001, 0.01, 0.1],
                    'depth': [4, 6, 8],
                    'l2_leaf_reg': [1, 3, 5, 7, 9],
                   }

# Realizando a busca aleatória
catboost_random = RandomizedSearchCV(estimator=catboost,
                                     param_distributions=param_grid_catboost,
                                     n_iter=50,
                                     random_state=42,
                                     cv=3,
                                     verbose=2,
                                     n_jobs=-1)

# Executando a busca aleatória
catboost_random.fit(X_train, y_train)

# Resultados da busca
print('Melhor score: %s' % catboost_random.best_score_)
print('Melhores hiperparâmetros: %s' % catboost_random.best_params_)

# Selecionando apenas os melhores resultados
catboost = CatBoostRegressor(**catboost_random.best_params_, random_state=42, verbose=False)
# Treinando o modelo otimizado
catboost.fit(X_train, y_train)

# Avaliando o desempenho do modelo nos dados de treinamento
pred_train_catboost = catboost.predict(X_train)
```

```
print('RMSE (treino):', np.sqrt(mean_squared_error(y_train, pred_train_catboost)))
print('R² (treino):', r2_score(y_train, pred_train_catboost))

# Avaliando o desempenho do modelo nos dados de teste
pred_test_catboost = catboost.predict(X_test)
print('RMSE (teste):', np.sqrt(mean_squared_error(y_test, pred_test_catboost)))
print('R² (teste):', r2_score(y_test, pred_test_catboost))

# Gráfico de dispersão: valor real vs valor predito
plt.figure(figsize=(15,10))
plt.scatter(y_test, pred_test_catboost, c='crimson')

# Determinar os limites dos eixos
p1 = max(max(pred_test_catboost), max(np.array(y_test)[0]))
p2 = min(min(pred_test_catboost), min(np.array(y_test)[0]))

# Plotar a linha diagonal
plt.plot([p1, p2], [p1, p2])

# Adicionar título e rótulos dos eixos
plt.title("Real vs Predito")
plt.xlabel('True Values', fontsize=15)
plt.ylabel('Predictions', fontsize=15)

# Ajustar proporções dos eixos
plt.axis('equal')

# Exibir o gráfico
plt.show()

Treinar o modelo CatBoost com os melhores hiperparâmetros encontrados
catboost = CatBoostRegressor(**catboost_random.best_params_, random_state=42, verbose=False)
catboost.fit(X_train, y_train)

# Calcular os valores SHAP
explainer = shap.Explainer(catboost)
shap_values = explainer.shap_values(X_test)

# Plotar o summary plot como gráfico de barras
shap.summary_plot(shap_values, X_test, plot_type='bar', color='hotpink', max_display=10)

# === Gráfico tipo 'bar' com escala personalizada para CatBoost ===
shap.summary_plot(
    shap_values,
    X_test,
    plot_type='bar',
    color='blue',
    max_display=10,
    show=False  # Impede que o gráfico apareça automaticamente
)

plt.xlim(0, 3.5)
plt.title("SHAP - CatBoost (Bar)")
plt.show()

import shap
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from catboost import CatBoostRegressor
```

```
# Treinar o modelo CatBoost com os melhores hiperparâmetros encontrados
catboost = CatBoostRegressor(**catboost_random.best_params_, random_state=42, verbose=False)
catboost.fit(X_train, y_train)

# Calcular os valores SHAP
explainer = shap.Explainer(catboost)
shap_values = explainer.shap_values(X_test)

# Calcular a média dos valores absolutos dos SHAP por variável
importance = np.abs(shap_values).mean(axis=0)

# Converter para porcentagem
importance_percent = 100 * importance / importance.sum()

# Ordenar por importância
sorted_indices = np.argsort(importance_percent)[::-1]
sorted_feature_names = X_train.columns[sorted_indices]
sorted_importance_percent = importance_percent[sorted_indices]

# Selecionar as 10 variáveis mais importantes
top_10_features = sorted_feature_names[:10]
top_10_importance = sorted_importance_percent[:10]

# Plotar gráfico manual com fonte ajustada
plt.figure(figsize=(10, 6))
plt.barh(top_10_features[::-1], top_10_importance[::-1], color='blue')
plt.xlabel('Importância (%)', fontsize=20)
#plt.ylabel('Variáveis', fontsize=14)
#plt.title('Importância das Variáveis - CatBoost (SHAP)', fontsize=16)
plt.xticks(fontsize=20)
plt.yticks(fontsize=20)
plt.xlim(0, 30)
plt.tight_layout()
plt.show()
```