PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS

ESCOLA POLITÉCNICA

CURSO DE GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO

GABRIEL TEIXEIRA ANDRADE SOUSA

PREDIÇÃO DE RESISTÊNCIA ANTIMICROBIANA EM PSEUDOMONAS AERUGINOSA COM APRENDIZAGEM DE MÁQUINA.

GOIÂNIA – GOIÁS

2022

GABRIEL TEIXEIRA ANDRADE SOUSA

PREDIÇÃO DE RESISTÊNCIA ANTIMICROBIANA EM PSEUDOMONAS AERUGINOSA COM APRENDIZAGEM DE MÁQUINA.

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia da Computação do Escola Politécnica da Pontifícia Universidade Católica de Goiás, como requisito parcial à obtenção do grau de bacharel em Engenharia da Computação.

Orientador: Clarimar José Coelho

Co-Orientador: Walcy Santos Rios

GOIÂNIA – GOIÁS

2022

GABRIEL TEIXEIRA ANDRADE SOUSA

PREDIÇÃO DE RESISTÊNCIA ANTIMICROBIANA EM PSEUDOMONAS AERUGINOSA COM APRENDIZAGEM DE MÁQUINA.

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia da Computação do Escola Politécnica da Pontifícia Universidade Católica de Goiás, como requisito parcial à obtenção do grau de bacharel em Engenharia da Computação.

Aprovada em: 08 de Dezembro de 2022

BANCA EXAMINADORA

Clarimar José Coelho (Orientador)
Escola Politécnica
Pontifícia Universidade Católica de Goiás - PUCGO

Walcy Santos Rios (Co-Orientador)
Centro de Excelência em Inteligência Artificial - CEIA
Universidade Federal de Goiás - UFG

Rafael Viana de Carvalho
Laboratório de Computação Científica - LCC

Douglas Vieira do Nascimento
Laboratório de Computação Científica - LCC
Universidade Federal de Goiás - UFG

AGRADECIMENTOS

Aos meus pais Paulo e Glauce, pelo amor, incentivo e apoio incondicional.

A minha irmã Paula, que mesmo de longe se fez presente durante todo o percurso.

A minha namorada Débora, que esteve comigo durante os momentos mais difíceis da produção desse trabalho e me deu forças pra seguir em frente

A meu caro amigo e co-orientador Walcy, pela preocupação e incentivo do meu desenvolvimento acadêmico.

A todos vocês, que foram essenciais para a conclusão desse último capítulo, a minha eterna gratidão. EU AMO VOCÊS!

“Nós só podemos ver um pouco do futuro, mas o suficiente para perceber que há muito a fazer”

(Alan Turing)

RESUMO

A queda na produção de novos antibióticos em conjunto com o aumento de bactérias resistentes, provou necessário encontrar formas alternativas de combater esse problema. A forma recomendada pela Organização Mundial de Saúde é a conscientização na administração dos antibióticos a pacientes em necessidade. Para esse fim, o aprendizado de máqunia tem se mostrado nos últimos anos como uma excelente solução. Com o objetivo de combater a bactéria multirresistente Pseudomonas aeruginosa, foi realizado a obtenção de amostras por meio do conhecido banco de dados público PATRIC. O tratamento dos dados das amostras por meio da análise k-mer se mostrou eficaz, porém computacionalmente custoso, limitando o escopo da pesquisa. A partir dai, foi realizado o treinamento dos algoritmos de aprendizagem de máquina Adaboost, Bagging, SVM, Árvores aleatórias, Regressão Logistica e GradientBoost. Como resultado, obteve-se um F1-score médio entre 0.7-0.8, evidenciando o potencial dessas ferramentas em auxiliar o tratamento de infecções bacterianas.

Palavras-chave: Aprendizagem de máquina. Resistência antimicrobiana. Pseudomonas Aeruginosa. Análise K-mer. Sequenciamento de Genoma Completo.

ABSTRACT

The drop in the production of new antibiotics along with the increase in resistant bacteria, proved necessary to find alternative ways to combat this problem. The recommendation given by the World Health Organization is the awareness in the administration of antibiotics to patients in need. In this aspect, machine learning has been shown in recent years as an excellent solution. With the objective of combating the multiresistant bacteria Pseudomonas aeruginosa, samples were obtained through the well-known public database PATRIC. Treatment of data through k-mer analysis proved to be effective, but computationally costly, limiting the scope of the research. Moving foward, machine learning algorithms training was carried out with Adaboost, Bagging, SVM, Random Trees, Logistic Regression and GradientBoost. The e F1-score averaging between 0.7-0.8, highlights the potential of these tools to help the treatment of bacterial infections.

Keywords: Machine Learning. Antimicrobial Resistance. Pseudomonas aeruginosa. K-mer Analysis. Whole Genome Sequencing.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

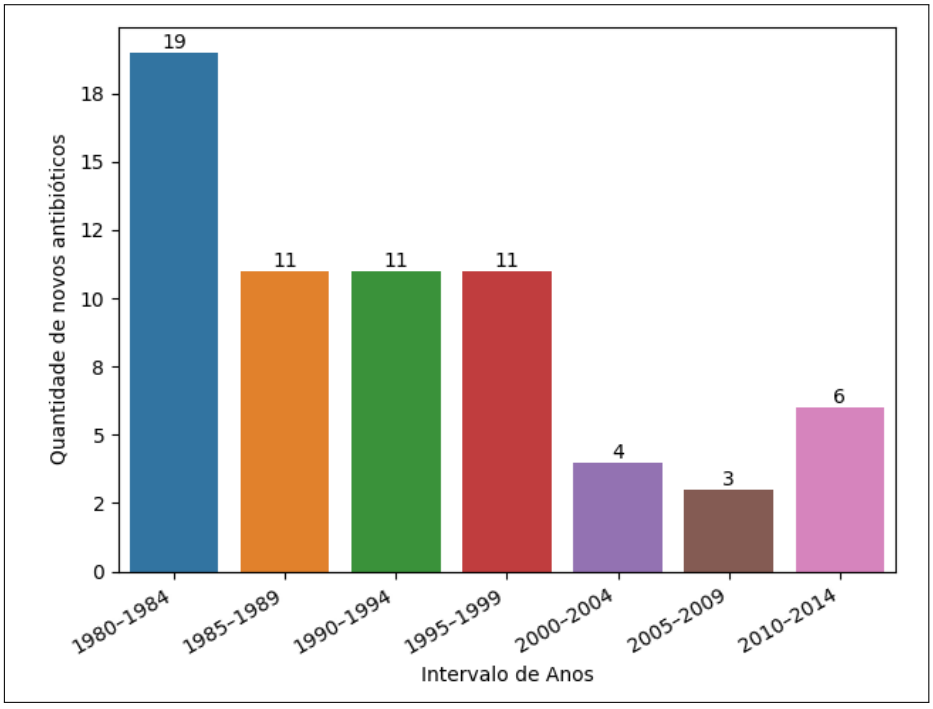
AMR	Resistência Antimicrobiana
ARG	Genes de Resistência Antimicrobiana
AST	Teste de Suscetibilidade Antimicrobiana
CAZ	Ceftazidima
CIP	Ciprofloxacino
CLSI	Instituto de Padrões Laboratoriais e Clínicos
CNN	Redes Neurais Convolucionais
DNA	Ácido Desoxirribonucleico
IA	Inteligência Artificial
LR	Regressão Logística
MAG	Máquina de Aumento de Gradiente
MEM	Meropenem
MIC	Concentração Inibitória Minima
NGS	Next-Generation Sequencing
OMS	Organização Mundial de Saúde
PCR	Reação em Cadeia da Polimerase
RF	Arvores Aleatórias
SNP	Polimorfismo de Nucleotídeo Único
SVC	Support Vector Classifier
SVM	Support Vector Machine
TOB	Tobramicina
WGS	Sequenciamento Completo do Genoma

SUMÁRIO

1 INTRODUÇÃO

Com a descoberta da penicilina por Alexander Fleming, em 1928, os antibióticos se tornaram peça chave no tratamento de infecções bacterianas. O descobrimento de mais antibióticos nos anos seguintes acabaram revolucionando todo o sistema de saúde. Mas, patógenos como as bactérias evoluiram e se tornaram resistentes. Nos últimos anos a situação tem piorado, uma vez que o uso de antibióticos aumentou e a velocidade do desenvolvimento de novos antibióticos diminuiu (VENTOLA, 2015). A Figura 1 representa o número de novos antibióticos ao longo dos anos.

Figura 1 – Número de novos antibióticos ao longo dos anos.



Fonte**:** The antibiotic resistance crisis: part 1: causes and threats (2015).

1.1 BACTÉRIA

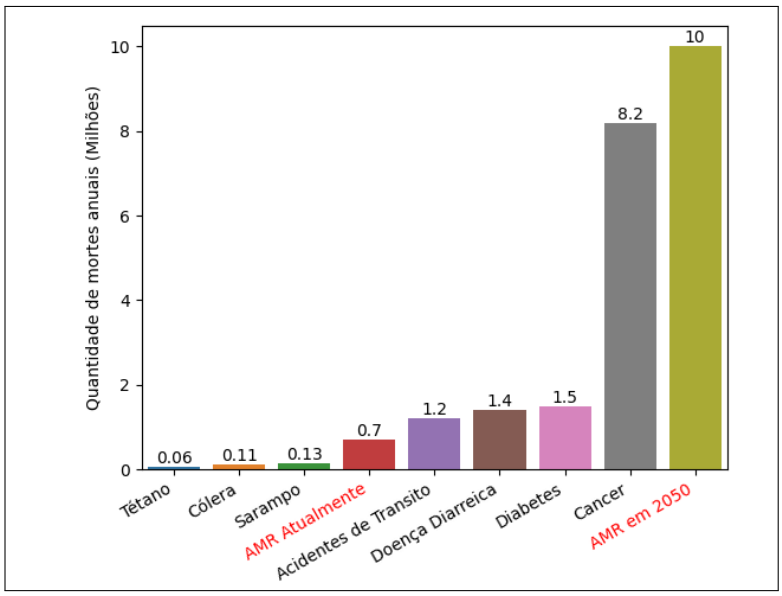
Normalmente pensados como hostís ou invasores que atacam nossos corpos, as bactérias, como outros organismos, estão apenas tentando viver e procriar. Viver as custas de um organismo hospedeiro é uma estratégia muito atrativa, e é possivel que todo organismo vivo na terra esteja sujeito a algum tipo de infecção. Bactérias são organismos unicelulares pequenos e estruturalmente simples quando comparados a vasta maioria das celulas eucarióticas. Dessa forma, seus genomas também podem ser considerados pequenos, sendo normalmente da ordem

de $1*10^6$ a $5*10^6$ pares de nucleotídeos, enquanto para os humanos esse numero é maior que $3*10^9$ (ALBERTS *et al.*, 2002).

A resistência a antibíoticos, ou Resistência Antimicrobiana (em inglês Antimicrobial Resistance (AMR)) se desenvolve quando a bactéria se adapta e cresce na presença dos mesmos. A resistência aos antibióticos é acelerada pelo uso indevido e excessivo de antibióticos, bem como pela má prevenção e controle de infecções. Como varios deles pertencem a mesma classe de medicamentos, resistência a um antibiótico específico pode levar a resistência de toda a classe relacionada. Quando desenvolvida em um único organismo ou local, a resistência pode se disseminar rapidamente e imprevisivelmente, como, por exemplo, quando uma bactéria troca material genético com outra (WHO, 2015).

O aumento da resistência antimicrobiana é uma das maiores ameaças à saúde em escala global, pois ela dificulta o uso convencional de antibióticos e eleva a taxa de ineficácia do tratamento antimicrobiano. O número anual de mortes em 2014 foi de aproximadamente 700 mil, mas estimativas indicam que em 2050 possa chegar a 10 milhões, podendo ter um custo global de 100 trilhões de dólares (O'NEILL, 2014).

Figura 2 – Principais causas de morte anualmente no mundo (e projeção).



Fonte**:** The Review on Antimicrobial Resistance Chaired by Jim O'Neill

Dentre as bactérias existentes, o maior risco se concentra naquelas resistente a mais de uma classe de antibióticos, também conhecidas como multirresistentes. Isso não só porque o tratamento delas é mais complicado devido a dificuldade de medicar o paciente, mas também por que elas tem a capacidade de passar essa multirresistência adiante para outras bactérias. A

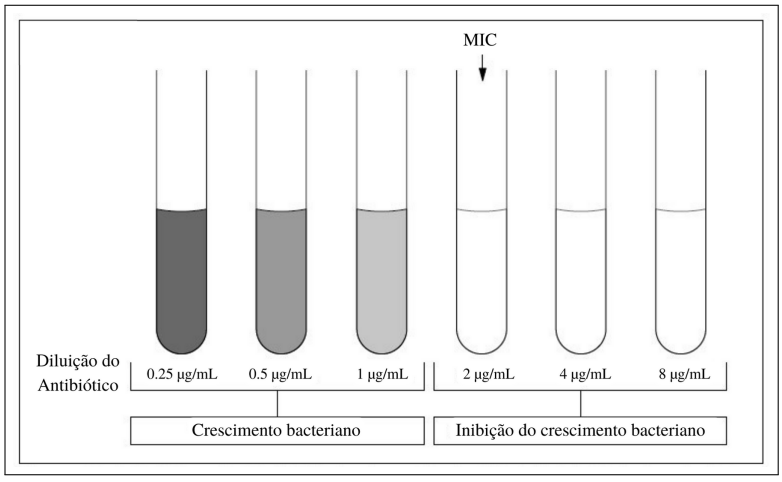
Organização Mundial de Saúde (OMS) divulgou uma lista (WHO, 2017) em 2017 das bactérias multirresistentes cuja necessidade de novos antibióticos é mais urgente, como apresenta a Figura 2. Dentre elas está a Pseudomonas Aeruginosa, um perigoso patógeno e um dos principais causadores de infecções hospitalares, e por isso, alvo deste estudo.

O Teste de Suscetibilidade Antimicrobiana (AST) é um procedimento utilizado para determinar quais antibióticos um organismo específico ou um grupo de organismos é suscetível ou não. Existem poucas atividades em um laboratório de microbiologia clínica que utilizem mais tempo e recursos do que o AST, avaliando a relevância quanto ao cuidado de pacientes com infecção ele é a atividade laboratorial mais importante. Esses testes são frequentemente utilizados para tratamento individual de pacientes, e os dados gerados a partir de AST servem para orientar Terapia Antimicrobiana Empírica (DOERN, 2011).

A Concentração Inibitória Minima (MIC) é a menor concentração (em µg/mL) que um antibiótico precisa para inibir o crescimento de uma cepa da bactéria testada. Um método quantitativo de AST, o MIC ajuda a determinar qual classe de antibiótico será mais efetiva, levando a um tratamento mais assertivo e que promove o combate contra AMR.

Após o AST, seguido do antibiótico está a interpretação de suscetibilidade: S (sensitível), I (intermediário), ou R (Resistente) e então o MIC (em µg/mL). Sensitivo implica que o organismo é inibido pela concentração da dose usual. Intermediário, implica que é inibido somente pela dose de concentração máxima recomendada. E resistente, implica que o organismo é resistente ao nível da dose permitida. Esses padrões de interpretação foram estabelecidos pelo Instituto de Padrões Laboratoriais e Clínicos (CLSI) (IDEXX, 2019), como apresenta a Figura 3.

Figura 3 – Teste de suscetibilidade a partir do MIC

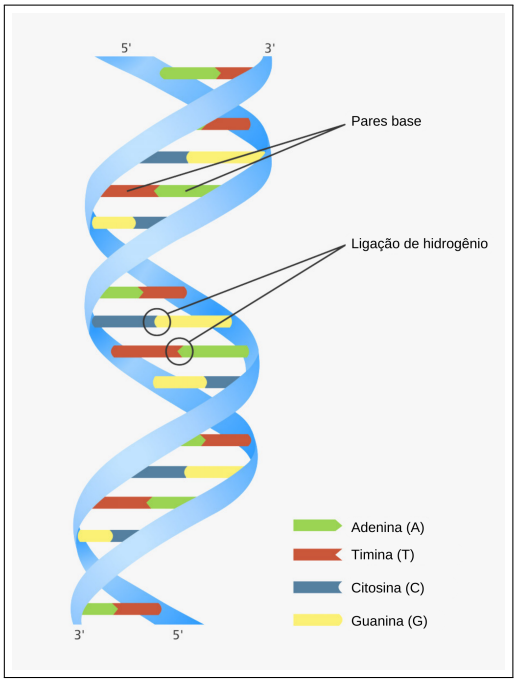


Fonte**:** Autoral

1.2 DNA

Para o estudos com inteligência artificial é fundamental o conhecimento do DNA, que será a principal fonte de dados. Descoberto em 1869 pelo suéco Friedrich Miescher, o Ácido Desoxirribonucleico (DNA) é um polímero de nucleotídeos que contém um código genético único. Os nucleotídeos são conhecidos popularmente pela base nitrogenada que os compõe, são elas a Adenina (A), Timina (T), Citosina (C) e Guanina (G). O DNA é uma molécula com duas fitas em formato de hélice dupla (Figura 4), cada fita é composta por longas sequências das bases mencionadas anteriormente. As bases em uma fita formam pares complementares (A com T e G com C) com a outra fita por meio de ligações de hidrogênio (BROWN, 2018).

Figura 4 – A molécula de DNA



Fonte**:** Genome Research Limited

O sequenciamento do DNA é uma forma de determinar a ordem das bases que compõe a molécula. A partir dessa sequência é possivel obter informações como quais pedaços do DNA contém genes e quais contém instruções regulatórias. Sequenciar um genoma completo (todo o DNA de um organismo) ainda é uma tarefa dificil. É uma atividade que requer quebrar o DNA em pedaços menores, sequenciar os pedaços e remontar o DNA formando uma única sequência. (NURK; KIRSCHE; RHIE, 2022).

As tecnologias de sequenciamento de DNA mais recentes são chamadas coletivamente de Next-Generation Sequencing(BUERMANS; DUNNEN, 2014). O conceito da NGS é

executar um grande número de pequenas reações de sequenciamento em paralelo. Dessa forma, grandes quantidades de DNA podem ser sequenciados com maior agilidade e menos custo. Por exemplo, em 2001 o custo do sequenciamento do genoma humano foi de quase 100 milhões de dólares (BUERMANS; DUNNEN, 2014), já em 2022 esse valor ficou em torno de 600 dólares (COLBY, 2022) .

Essa facilitação no sequenciamento genético fez com que uma imensa quantidade de dados fossem gerados nos últimos anos, e com isso o surgimento de bancos de dados para armazenar essas informações. Esses bancos são utilizados para diversas áreas de estudo, inclusive, na Inteligência Artificial (IA), o sucesso desses métodos depende da abrangência e qualidade dos dados, que se tratando de AMR consistem geralmente em Genes de Resistência Antimicrobiana (ARG) e de resultados de AST. Os principais bancos de dados baseados em AMR foram elencados e mostrados na Tabela 1.

Tabela 1 – Bancos de dados típicos de AMR

Nome	Descrição	Site
PATRIC	Genomas bacterianos com fenótipos AMR e MIC	https://patricbrc.org/
CARD	ARG e mecanismos de resistência	https://card.mcmaster.ca/
ARDB	Informações de ARG	https://ardb.cbcb.umd.edu/
BacMet	Biocida antibacteriano e genes de resistência metal	http://bacmet.biomedicine.gu.se
ARG-ANNOT	ARG e mutações pontuais	http://www.mediterranee-infection.com/

Fonte**:** A review of artificial intelligence applications for antimicrobial resistance

1.3 APLICAÇÕES DE INTELIGÊNCIA ARTIFICIAL PARA AMR

O estudo de identificação por culturas ou Reação em Cadeia da Polimerase (PCR) presentes no AST para a identificação de AMR estão sendo substituidas pelo grande fluxo de dados metagenômicos (Estudo da estrutura e função da sequência de nucleotideos para todos os organismos). Essa mudança, juntamente com o crescimento do número de microorganismos resistentes tem gerado dados complexos e de larga escala, implicando no aumento do uso de aprendizado de máquina, que é uma ótima ferramenta para analizar os conjuntos de dados diversos e fragmentados de AMR. Dessa forma, explorando os padrões, avaliação e predição de resistência, cria-se um precedente para que novas políticas quanto ao uso de drogas antibióticas possam ser feitas futuramente.

1.3.1 Predição de AMR

Existem atualmente dois métodos que se destacam na realização do diagnóstico de AMR, o anteriormente mencionado AST (Ao menos 24h) que não é eficiente e não explica os mecanismos de AMR, e o Sequenciamento de Genoma completo para AST (WGS-AST). Este último método fornece um diagnóstico rápido e preciso, porém requer um conjunto de dados grande e multidimensional para extrair informações de forma efetiva. É nesse ponto que é possivel perceber a grande contribuição da aprendizagem de máquina, visto que acombinação desses métodos consegue obter bons resultados em menos de 3 horas (LV; DENG; ZHANG, 2020).

(YANG *et al.*, 2018) e (KOUCHAKI *et al.*, 2019) analisaram AMR usando diferentes algoritmos de aprendizagem de máquina (Support Vector Machine (SVM), Regressão Logística (LR) e Arvores Aleatórias (RF)), treinados com Sequenciamento Completo do Genoma (WGS) atingiram uma acurácia alta na predição de AMR. Algoritmos de aprendizagem profunda também mostraram um potencial significativo na predição de novos antibióticos, genes AMR, e peptídeos de AMR (ARANGO *et al.*, 2018), (STOKES *et al.*, 2020), (VELTRO; ET, 2018). Contúdo, esses estudos focaram em variantes dos genomas (como, Polimorfismo de Nucleotídeo Único (SNPs)) ou outras características relacionadas a genes de resistência identificados em estudos anteriores ou bancos de dados de resistência.

No patógeno Pseudomonas Aeruginosa, mesmo informações da sequência genômica são insuficientes para predizer AMR em todas as amostras clínicas, devido a sua multirresistência e por ter um elevado índice de mutação (KOS *et al.*, 2015), se mostrando eficiente apenas para certos antibióticos. Dessa forma, o potencial dos modelos de aprendizagem de máquina para predição de AMR sem a utilização de bases de dados de mutação de resistências conhecida ou de genes anotados requer maior esclarecimento.

1.4 OBJETIVOS

1.4.1 Objetivo Geral

O objetivo deste trabalho é realizar a predição do fenótipo de resistência antimicrobiana em amostras da bactéria Pseudomonas Aeruginosa, a partir de seu sequenciamento do genoma completo, utilizando algoritmos de aprendizagem de máquina de classificação.

1.4.2 Objetivos Específicos

Os objetivos específicos deste trabalho são:

a) Avaliar a eficácia e viabilidade de algoritmos de aprendizagem (Adaboost, Bagging, SVM, Árvores aleatórias, Regressão Logistica, GradientBoost) na classificação do fenótipo de resistência da bactéria Pseudomonas Aeruginosa.

b) Avaliar o desempenho dos algoritmos (Adaboost, Bagging, SVM, Árvores aleatórias, Regressão Logistica, GradientBoost) durante o treinamento com o WGS.

c) Propor métodos para melhorar a predição de AMR em bactérias multirresistentes.

2 MATERIAL E MÉTODOS

2.1 FUNDAMENTAÇÃO TEÓRICA

Os trabalhos apresentados a seguir, foram os que forneceram maior fundamentação teórica para esta monografia.

2.1.1 Prevendo resistência antimicrobiana em Pseudomonas aeruginosa com aprendizagem de máquina auxiliado por diagnóstico molecular

Nesse trabalho, os autores utilizaram 414 amostras de Pseudomonas Aeruginosa coletadas, e com elas foi realizado o sequenciamento genético, o teste de suscetibilidade antimicrobiana, além de outros métodos de identificação de genes de resistência. Com isso, integraram dados genômicos, transcriptomicos e fenotípicos em perfis de resistência antibiótica das amostras.

Posteriormente, foi realizada uma abordagem com aprendizagem de máquina para identificar conjuntos de marcadores moleculares que permitiam uma predição de AMR confiável em quatro classes de antibióticos. Dessa forma, a utilização de informações sobre a presença ou ausência de genes, variações de sequência genética entre genes e perfis isolados ou combinados de expressão genética, resultaram em predições de sensibilidade alta (0.8-0.9) ou muito alta (> 0.9), com o algoritmo de aprendizagem de máquina SVM (KHALEDI; WEIMANN; SCHNIEDERJANS, 2020).

A informação de expressão genética melhorou a performance da predição de resistência para todas as drogas, exceto ciprofloxacina. Os resultados sugerem que uma ferramenta de classificação de resistência baseada e markadores genômicos e transcriptômicos podem melhorar a acurácia da testagem de resistência antimicrobiana. No trabalho é ressaltado a importância que os dados transcriptômicos tiveram nos resultados das predições quando comparados com a informação genômica sozinha. Os marcadores genéticos encontrados incluiam determinantes de AMR já conhecidos (como os genes: gyrA, ampC, oprD e bombas de efluxo), assim como novos marcadores que ainda não haviam sido associados a AMR.

2.1.2 Aprendizagem de máquina para resistência antimicrobiana

(SANTERRE *et al.*, 2016) O estudo se concentrou no uso de aprendizado de máquina para prever fenótipos de AMR, que se refere à capacidade de um microrganismo (como bactérias) de sobreviver à exposição a um agente antimicrobiano (como um antibiótico). Os autores usaram

um algoritmo de classificação de floresta aleatória (RF) para analisar os dados e obtiveram precisões relativamente altas, chegando a 92%.

Eles descobriram que a precisão do algoritmo diminuiu à medida que o número de isolados usados para treinamento aumentava, mas acabou se estabilizando em tamanhos de amostra maiores. Isso significa que, até certo ponto, usar mais dados de treinamento pode melhorar a precisão do modelo, mas além desse ponto, adicionar mais dados não melhora significativamente o desempenho do modelo.

Os autores também discutiram os objetivos potenciais do uso de aprendizado de máquina para estudar AMR, incluindo maximizar a precisão da classificação, maximizar a precisão da generalização e agregar a seleção de recursos. Maximizar a precisão da classificação refere-se a alcançar a maior precisão possível na previsão de fenótipos de AMR. Maximizar a precisão da generalização refere-se à capacidade do modelo de prever com precisão fenótipos de AMR em diferentes bactérias ou em diferentes antibióticos. Agregar a seleção de recursos envolve identificar os recursos mais importantes (como mutações genéticas específicas) que contribuem para a AMR e agrupá-los em grupos significativos.

2.1.3 Predição de resistência antimicrobiana baseado em sequenciamento completo do genoma e aprendizagem de máquina

Nesse estudo, os autores avaliaram os seguintes algoritmos de aprendizagem de máquina: regressão logística(RL), Suport Vector Machine (SVM), Random Forest (RF), e Redes Neurais Convolucionais (CNN) para a predição de AMR em cepas da bacéria Escherichia Coli contra os antibióticos ciprofloxacino, cefotaxima, ceftazidima e gentamicina. Para a codificação dos dados WGS utilizados, foi utilizado 3 métodos diferentes, são eles: label encoding, one-hot encoding e FCGR encoding. Os autores demonstraram que esses modelos conseguiram efetivamente predizer os fenótipos AMR, em geral RFs e CNNs performaram melhor do que LR e SVM com a métrica AUC por volta de 0.96 (REN; CHAKRABORTY; DOIJAD, 2022).

2.2 DADOS GENÉTICOS E FENÓTIPOS AMR

Uma vez definido qual seria o organismo estudado, neste caso o patógeno Pseudomonas aeruginosa, os próximos passos para obter os dados necessários seriam: realizar a coleta das amostras; o sequenciamento genético por meio de algum método NGS; e para cada amostra realizar o AST por meio do MIC para obter os fenótipos AMR (sensível, resistente

ou intermediário). Só após concluir esses passos é possível começar o tratamento dos dados para prepara-los para os algoritmos de aprendizagem de máquina. Devido à escasses de tempo, recursos, e mão de obra qualificada, a melhor forma de conseguir esses dados foi buscando em bancos de dados genéticos públicos, como os mencionados anteriormente (Tabela 1), devido a sua grande variedade de dados genômicos, AST e MIC, o banco do qual os dados deste estudo foram retirados é o PATRIC (WATTAM *et al.*, 2017a) (WATTAM *et al.*, 2017b). Mais especificamente o estudo mencionado em fundamentação teórica (KHALEDI; WEIMANN; SCHNIEDERJANS, 2020) e identificado pelo PMID 32048461.

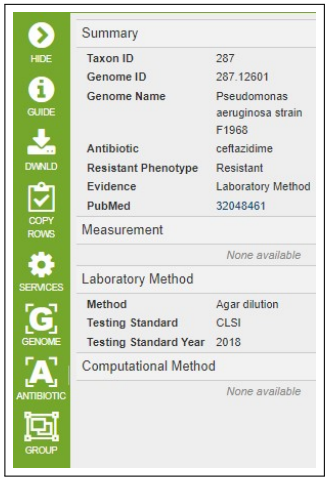
O Acesso aos dados de fenótipos AMR é realizado como indicado na Figura 5, onde é especificado o nome da cepa da bactéria, o antibiótico, o fenótipo de resistência e o número PMID do estudo. Já o acesso ao WGS é feito a partir do sumário das cepas, no botão de subtitulo ”Genome” (Figura 6).

Figura 5 – Lista de cepas e fenótipos de resistência no PATRIC

Genome Name	Antibiotic	Resistant Phenotype	Pubmed
Pseudomonas aeruginosa strain F1968	ceftazidime	Resistant	32048461
Pseudomonas aeruginosa strain CH2674	ceftazidime	Intermediate	32048461
Pseudomonas aeruginosa strain CH4990	tobramycin	Susceptible	32048461
Pseudomonas aeruginosa strain CH5262	ceftazidime	Susceptible	32048461
Pseudomonas aeruginosa strain ESP083	ciprofloxacin	Resistant	32048461
Pseudomonas aeruginosa strain CH4411	ciprofloxacin	Resistant	32048461
Pseudomonas aeruginosa strain CH4860	tobramycin	Susceptible	32048461
Pseudomonas aeruginosa strain CH4446	ceftazidime	Susceptible	32048461

Fonte**:** https://www.bv-brc.org/view/Bacteria.2022.

Figura 6 – Sumário dos dados de uma cepa no PATRIC

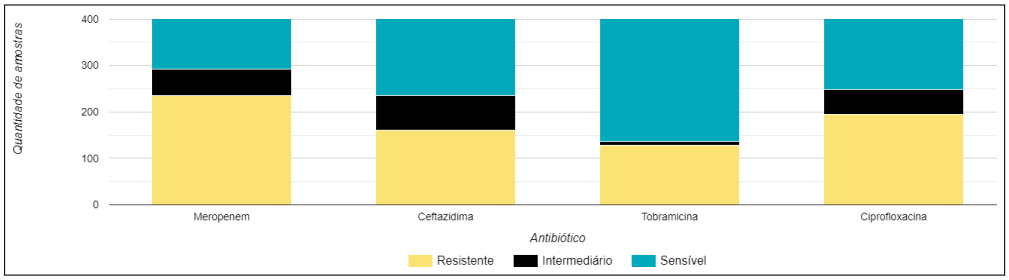


Fonte**:** https://www.bv-brc.org/view/Bacteria.2022.

Nesse estudo, foram coletadas 414 amostras de P.aeruginosa de variados locais da

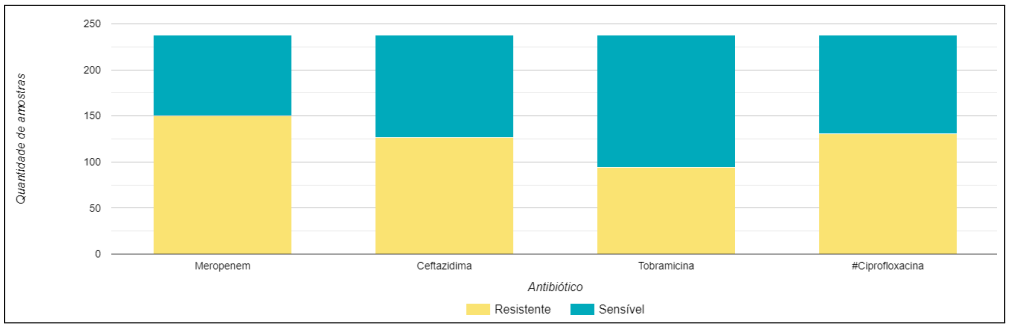
Europa, tanto de universidades quanto de clínicas. Todos as amostras tiveram sua suscetibilidade testada para os quatro antibióticos anti-pseudomonas mais comuns, sendo eles: Tobramicina (TOB), Ciprofloxacino (CIP), Meropenem (MEM), and Ceftazidima (CAZ). O teste de concentração mínima inibitória foi aplicado de forma triplicada para cada amostra, se os resultados variassem, até cinco réplicas eram utilizadas. Apenas amostras com ao menos três resultados compatíveis foram incluídas no estudo. A maioria das amostras foram categorizadas como multirresistentes, ou seja, resistente a mais de três classes de antibióticos. Com o objetivo de utilizar apenas algoritmos de aprendizagem de máquina de classificação, todas amostras que possuiam fenótipo intermediário para ao menos um dos antibióticos foram retiradas do estudo, passando assim de 414 para 238 amostras. As proporções dos fenótipos para cada antibiótico esta representado na Figura 7 e Figura 8 mostrando o antes e depois.

Figura 7 – Fenótipos por antibiótico (com intermediário)



Fonte**:** Autoral.

Figura 8 – Fenótipos por antibiótico (sem intermediário)



Fonte**:** Autoral.

Os dados do WGS armazenados no PATRIC estão no formato FASTA (ou FNA), é um arquivo em formato de texto que representa tanto sequências nucleotídicas quanto peptídicas. Uma sequência em FASTA começa com uma descrição em uma única linha, seguido pelas linhas que formam a sequência (Figura 9). A linha de descrição pode ser diferênciada da sequência genética pelo simbolo ”>” na primeira coluna. É recomendado que cada linha de texto possua

menos que 80 caracteres de comprimento (ZHANG, 2018). Os arquivos fasta obtidos possuiam uma média de 7 megabytes, com um comprimento total médio de 7.4 milhões de caracteres.

Figura 9 – Modelo de arquivo FASTA

```
>NC_017548.1 Pseudomonas aeruginosa M18, complete sequence

TTTAAAGAGACCGGCGATTCTAGTGAAATCGAACGGGCAGGTCAATTTCCAACCAG
CGATGACGTAATAGATAGATACAAGGAAGTCATTTTTCTTTTAAAGGATAGAAACG
GTTAATGCTCTTGGGACGGCGCTTTTCTGTGCATAACTCGATGAAGCCCAGCAA. . .
```

Fonte**:** Autoral.

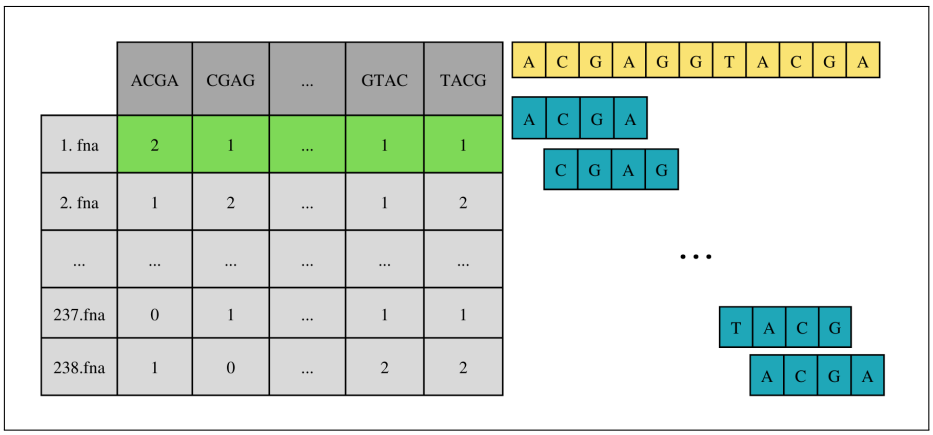
2.3 TRATAMENTO DOS DADOS

Os dados dos arquivos FASTA, são enormes sequências de caracteres das quatro bases nucleotídicas (ACGT), por isso, para preparar os dados para servirem de entrada a algoritmos de aprendizagem de máquina é preciso utilizar de técnicas de codificação e assim conseguir valores numéricos. Para alcançar esse objetivo, foi utilizado a análise k-mer, um método bem disseminado em áreas da bioinformática como a montagem de genomas e transcriptomas, sequenciamento metagenômico e para correção de erros de leitura do sequenciamento (ONATE; BATTO; JUSTE, 2015).

K-mer é uma sequência de caracteres de tamanho k em um texto (em amarelo) (ou de nucleotídeos em uma sequência de DNA). Para obter todos os k-mer de uma sequência é preciso selecionar a primeira janela de tamanho k e mover apenas um caracter para o inicio do proximo k-mer, também conhecido como janela deslizante (em azul), eliminando a influência de um ponto inicial arbitrário (GUNASEKARAN; RAMALAKSHMI; AROKIARAJ, 2021). O resultado da análise k-mer é a formação de um dicionário das frequências (em cinza) de k-mer que pode ser utilizado para treinamento de modelos de inteligência artificial (Figura 10).

Porém, a formação do dicionário é um processo com grande custo computacional. Nesse estudo foi utilizado um computador com processador Intel I5-6500 de 4 núcleos e 3.60 GHz de frequência, e 16 gigabytes de memória ram. Com essas especificações foi possível realizar a análise k-mer até o valor k=9, limitado pela memória ram. A Tabela 2 apresenta as características dos dicionários gerados.

Figura 10 – Análise K-mer. K=4.



Fonte**:** Autoral.

Tabela 2 – Características dos dicionários k-mer.

	Tamanho do Arquivo (Mb)	Quantidade de Colunas
k = 5	2.007	1025
k = 6	7.034	4097
k = 7	23.846	16385
k = 8	78.934	65537
k = 9	250.214	262145

Fonte**:** Autoral

2.4 APRENDIZAGEM DE MÁQUINA

Popularmente conhecida como o campo de estudo que permite que computadores aprendam sem ser explicitamente programado, a aprendizagem de máquina depende de diferentes algoritmos para solucionar problemas a partir informações pré-existente. Cientistas de dados ressaltam que não existe um único algoritmo que age como uma bala de prata e é capaz de resolver todos os tipos de problemas. O tipo de algoritmo utilizado depende do problema que se quer resolver, o número de variáveis, o modelo que melhor se encaixa, etc (MAHESH, 2018).

2.4.1 Aprendizado Supervisionado

A Aprendizagem supervisionada é a tarefa de aprender a função que mapeia uma entrada (E) para uma saida (S), baseado nos exemplos E/S anteriormente fornecidos. Dessa forma, o algoritmo irá montar uma função a partir dos dados de treinamento categorizados de um conjunto de treinamento, ou seja, a aprendizagem supervisionada requer assistência externa para conseguir esses dados categorizados. O conjunto de entrada do algoritmo é dividido em conjunto de treinamento e teste, onde o de treino possui uma variável de saida que precisa ser

classificada ou predita (RAY, 2019).

2.4.2 Aprendendo a prever Resistência

Na aplicação da aprendizagem de máquina neste trabalho, foi utilizado o aprendizado supervisionado, uma vez que a entrada dos algoritmos gerada a partir da análise k-mer (Fig) utiliza da categorização fornecida pelo AST, ou seja, utiliza os fenótipos de resistência aos antibióticos testados. Dessa forma, o conjunto de treinamento utiliza dos fenótipos AMR para aprender a classificar o novo WGS como resistente ou sensível. Justificando assim a escolha de somente algoritmos de classificação nos treinamentos e testes realizados.

2.5 ALGORITMOS DE CLASSIFICAÇÃO

Os algoritmos de classificação abaixo foram escolhidos com o objetivo de avaliar a viabilidade do aprendizado de máquina aplicado ao WGS por meio da análise k-mer.

2.5.1 Classificador Adaboost

O algoritmo de classificação Adaboost funciona pegando um conjunto fraco, nesse caso os k-mers obtidos pelos genomas, e os classifica a partir de um processo iterativo de refinamento. Esse conjunto de k-mer distintos e seus respectivos pesos passam a ser classificadores que são usados para predizer o fenótipo de um novo genoma (TSAI; HUNG, 2021) como mostrado na Figura 11.

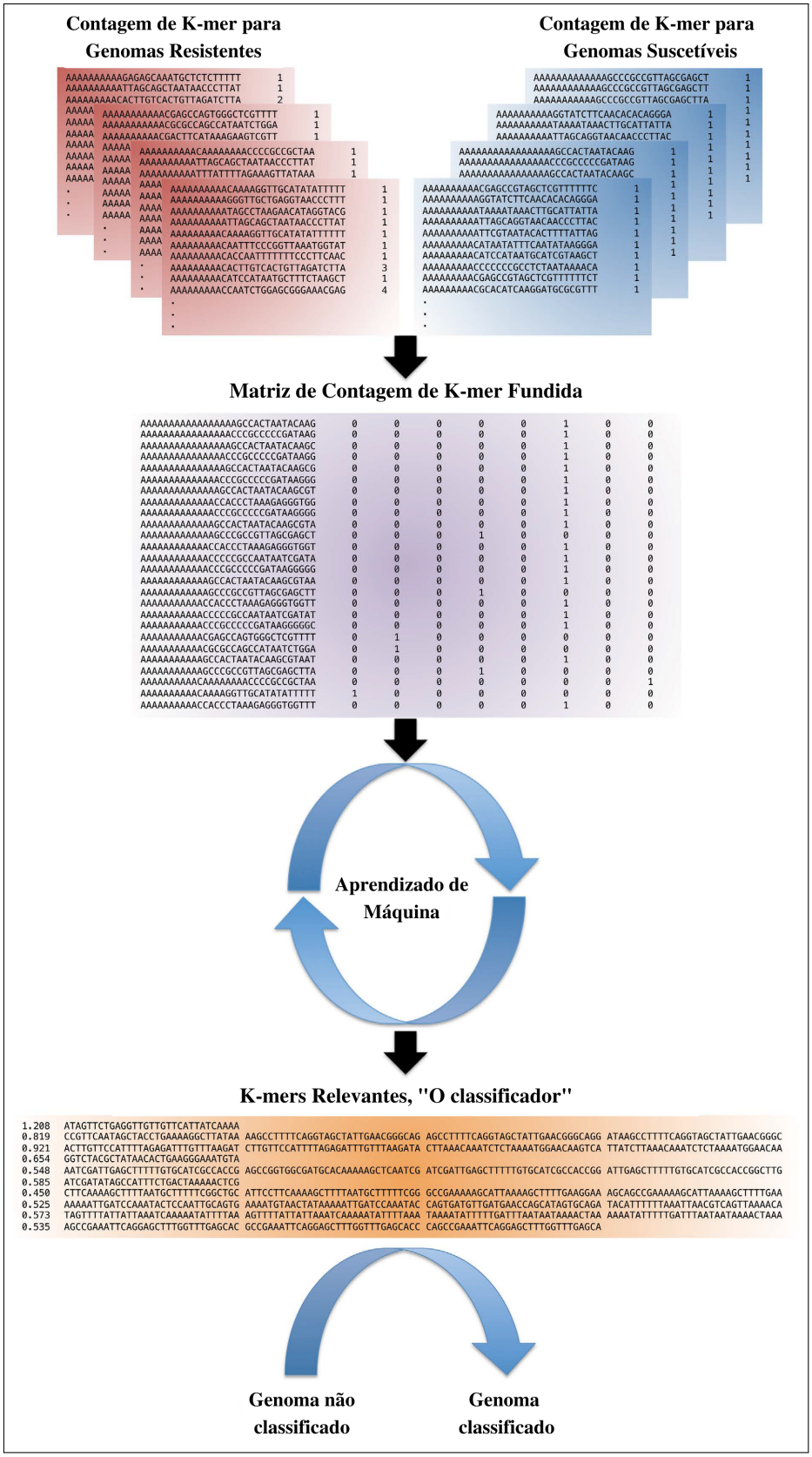
Os parâmetros utilizados no treinamento do classificador Adaboost foram: Número de estimadores = 50; Taxa de aprendizado = 1.0; algorítmo = SAMME.R.

2.5.2 Classificador de Bagging

Um classificador de Bagging funciona da seguinte forma, dado um conjunto de dados com m amostras, é selecionado uma amostra aleatóriamente e então é copiada para o cojunto de amostragem, essa é mantida no conjunto original podendo ser escolhida novamente na próxima seleção. Repetindo o processo ”m” vezes resulta em um conjunto de dados de tamanho ”m” onde algumas amostras do conjunto original podem aparecer mais de uma vez ou nenhuma vez. Aplicando esse processo por ”T” vezes produz ”T” conjuntos de dados com ”m” amostras.

A previsão por bagging é um método onde multiplas versões de um preditor são

Figura 11 – Classificador Adaboost



Fonte**:** (TSAI; HUNG, 2021)

treinados nos ”T” conjuntos de dados para então formar um único preditor agregado, que a partir dos votos dessas versões realiza a previsão de uma classe (BREIMAN, 1996).

Os parâmetros utilizados no treinamento do classificador Bagging foram: Número de estimadores = 10; Máximo de amostras = 1.0; Máximo de características = 1.0.

2.5.3 Classificador de Aumento de Gradiente

A Máquina de Aumento de Gradiente (MAG) está entre os algoritmos de aprendizagem de máquina de conjuntos (ensemble) para problemas de regressão e de classificação. Na MAG, é atribuído peso aos dados, então o procedimento de aprendizagem e um modelo para treinar os dados a partir de uma combinação de um conjunto fraco é construido. Então o algoritmo é reconfigurado iterativamente para averiguar se ele pode performar melhor com outro conjunto de parâmetros (AZIZ; AKHIR; AZIZ, 2020).

Os parâmetros utilizados no treinamento do classificador de Aumento de Gradiente foram: Taxa de aprendizado = 0.1; Número de estimadores = 100; Sub-amostragem = 1.0; critério = friedman-mse; divisão mínima = 2; Folhas mínimas = 1; Profundidade máxima = 3.

2.5.4 Regressão Logistica

A regressão logística é um modelo estatístico utilizado para realizar uma classificação binária. Ele mostra a relação entre os recursos e, em seguida, calcula a probabilidade de um determinado resultado. A regressão logística é usada no aprendizado de máquina para ajudar a criar previsões precisas.

Na regressão logística, a variável dependente é uma variável binária que pode assumir um de dois valores possíveis, como "sim"ou "não", "verdadeiro"ou "falso"ou "0"ou "1". As variáveis independentes, também conhecidas como preditores ou recursos, são as variáveis usadas para prever o valor da variável dependente.

O objetivo da regressão logística é encontrar o melhor modelo de ajuste que possa prever a probabilidade da variável dependente com base nas variáveis independentes. Isso é feito ajustando uma curva logística aos dados e estimando a probabilidade da variável dependente com base nos valores das variáveis independentes.(FERNANDES; FILHO; ROCHA, 2020).

Os parâmetros utilizados no treinamento do classificador de Regressão Logistica foram: norma de penalidade = l2; Tolerância =0.0001; C = 1.0; Algoritmo = lbfgs; Máximo de iterações = 100.

2.5.5 Classificador de Florestas Aleatórias

As florestas aleatórias (popularmente conhecida pelo seu nome em inglês ”Random Forest” (RF)), é uma extensão do Bagging, onde uma seleção aleatória de características é introduzida por cima do bagging. Especificamente, arvores de decisão tradicionais selecionam uma divisão ótima de características dentre um conjunto em cada nó, já a RF seleciona em um subconjunto de n características gerado aleatóriamente para cada nó (ZHOU, 2021).

Em um classificador Random Forest, várias árvores de decisão são treinadas em diferentes subconjuntos dos dados de treinamento. Durante o processo de treinamento, cada árvore de decisão é construída usando um subconjunto diferente dos dados de treinamento e um subconjunto diferente dos recursos. A previsão final é feita pela média das previsões de todas as árvores de decisão individuais.

Os parâmetros utilizados no treinamento do classificador de Florestas Aleatórias foram: Número de estimadores = 100; Critério = gini; Mínimo de divisões = 2; minimo de folhas = 1; Máximo de características = sqrt.

2.5.6 SVM (SVC)

A máquina de vetor de suporte (em inglês Support Vector Machine (SVM)) é um algoritmo de aprendizado de máquina supervisionado que pode ser usado para desafios de classificação ou regressão. Seu foco maior é no treinamento e classificação de um conjunto de dados (Support Vector Classifier (SVC)) . Nesse algoritmo, cada item de dados é plotado como um ponto no espaço n-dimensional (onde n é o número de recursos que você tem), com o valor de cada recurso sendo o valor de uma determinada coordenada. Então, é executada a classificação encontrando o hiperplano que melhor diferencia as duas classes (CHANG; LIN, 2022).

Os parâmetros utilizados no treinamento do classificador SVM foram: C = 1.0; Kernel = rbf; gamma = scale; Tolerância = 0.001; Tamanho de cache = 200.

2.6 MÉTRICAS PARA AVALIAR ALGORITMOS DE CLASSIFICAÇÃO

2.6.1 Precisão

A precisão explica quantos dos casos previstos realmente eram positivos. É comumente utilizada em casos onde os falsos positivos são uma preocupação maior que os falsos

negativos, e é definido como o número de verdadeiros positivos dividido pelo número de positivos previstos.

$$Precisao = \frac{VerdadeiroPositivo}{VerdadeiroPositivo+FalsoPositivo}$$

2.6.2 Sensibilidade (Recall)

A sensibilidade explica quanto dos casos positivos realmente foi possivel prever corretamente. É comumente utilizada em casos em que falso negativo é de preocupação maior do que falso positivo, e é definido como o número de positivos verdadeiros dividido pelo número total de positivos reais.

$$Sensibilidade = \frac{VerdadeiroPositivo}{VerdadeiroPositivo+FalsoNegativo}$$

2.6.3 F1-Score

A métrica F1 fornece uma combinação das métricas de precisão e sensibilidiade, e tem valor máximo quando precisão é igual a sensibilidade. O F1-Score pune casos de valores extremos e é definido pela média harmonica da precisão e da sensibilidade. Por isso, foi escolhida como a principal métrica de avaliação dos modelos de aprendizagem.

$$F1\text{-}Score = 2 * \frac{Precisao*Sensibilidade}{Preciso+Sensibilidade}$$

2.6.4 Curva AUC-ROC

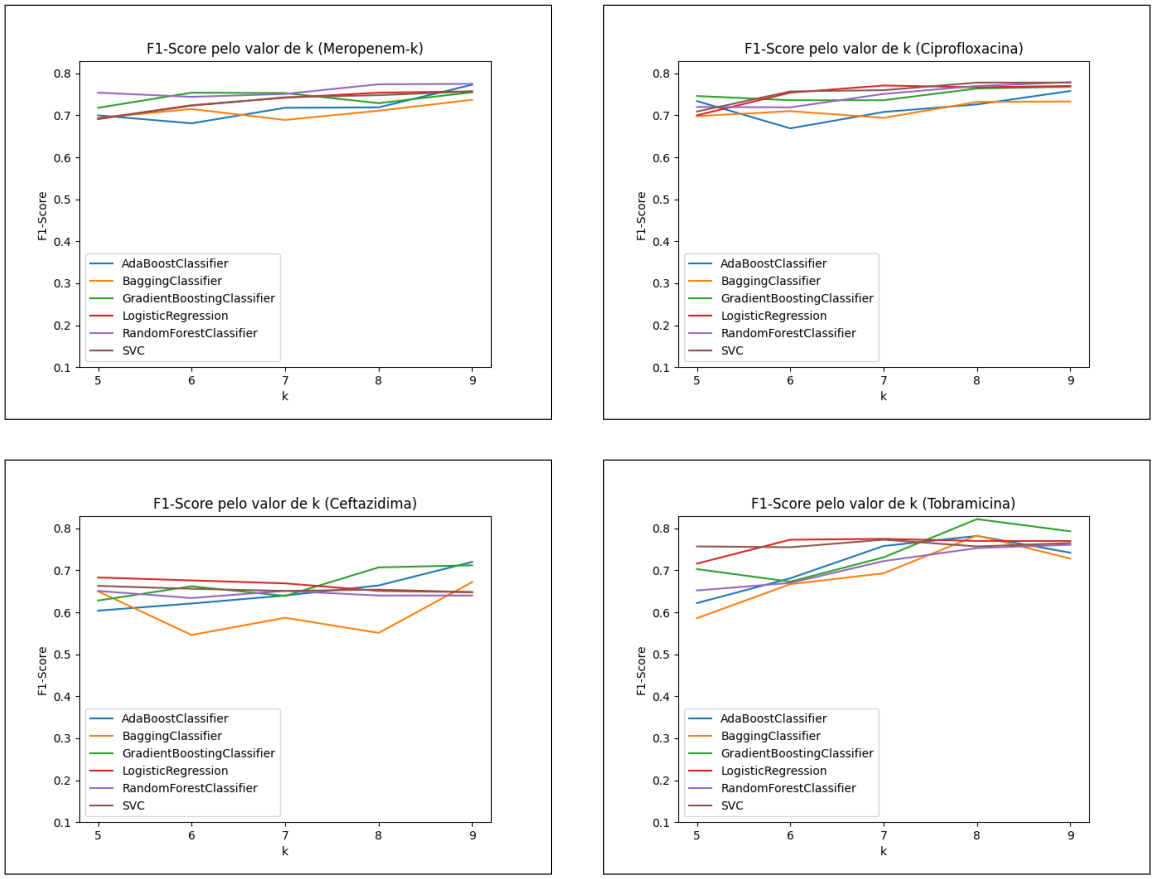
A curva ROC (Receiver Operator Characteristic) é uma curva de probabilidade que traça a taxa dos verdadeiros positivos pela taxa dos falsos positivos. A área embaixo da curva (AUC) é a medida da habilidade do classificador de distinguir entre as classes. Quanto maior o AUC melhor é a performance do modelo, de forma simplificada, quando AUC é igual a 1 significa que o classificador foi capaz de distinguir perfeitamente entre as classes positivas e negativas, e quando AUC é igual a 0 o classificador preveria todos os positivos como falsos e vice e versa.

3 RESULTADOS

Os treinamentos foram realizados utilizando a biblioteca scikit-learn. Dessa forma, foi utilizado uma proporção de treinamento/teste de 80/20 e também a técnica de validação cruzada com uma divisão de 10 subgrupos para garantir a generalização dos modelos.

O primeiro resultado a ser analisado é o valor do F1-score de acordo com o tamanho do k-mer, a Figura 13 mostra o aumento de k, para cada um dos antibióticos testados. O valor F1 demonstrou aumento gradual, e com k=9 obteve-se valores entre 0.7 e 0.8, com exceção do antibiótico Ceftazidima, que ficou entre 0.6 e 0.7. Os modelos que obtiveram melhor desempenho foram Adaboost e GradientBoost (Tabela 3) com médias de F1-score de 0.75 e 0.76 respectivamente, e desvio padrão < 0.1.

Figura 12 – F1-score x k-mer.



Fonte**:** Autoral.

O aumento do desempenho com k=9 gera expectativa que para valores maiores o valor de F1-Score também aumente. Outros estudos semelhantes geraram precedentes que corrobora com essa hipótese (KHALEDI; WEIMANN; SCHNIEDERJANS, 2020) (REN; CHAKRABORTY; DOIJAD, 2022).

Tabela 3 – F1-score para k=9.

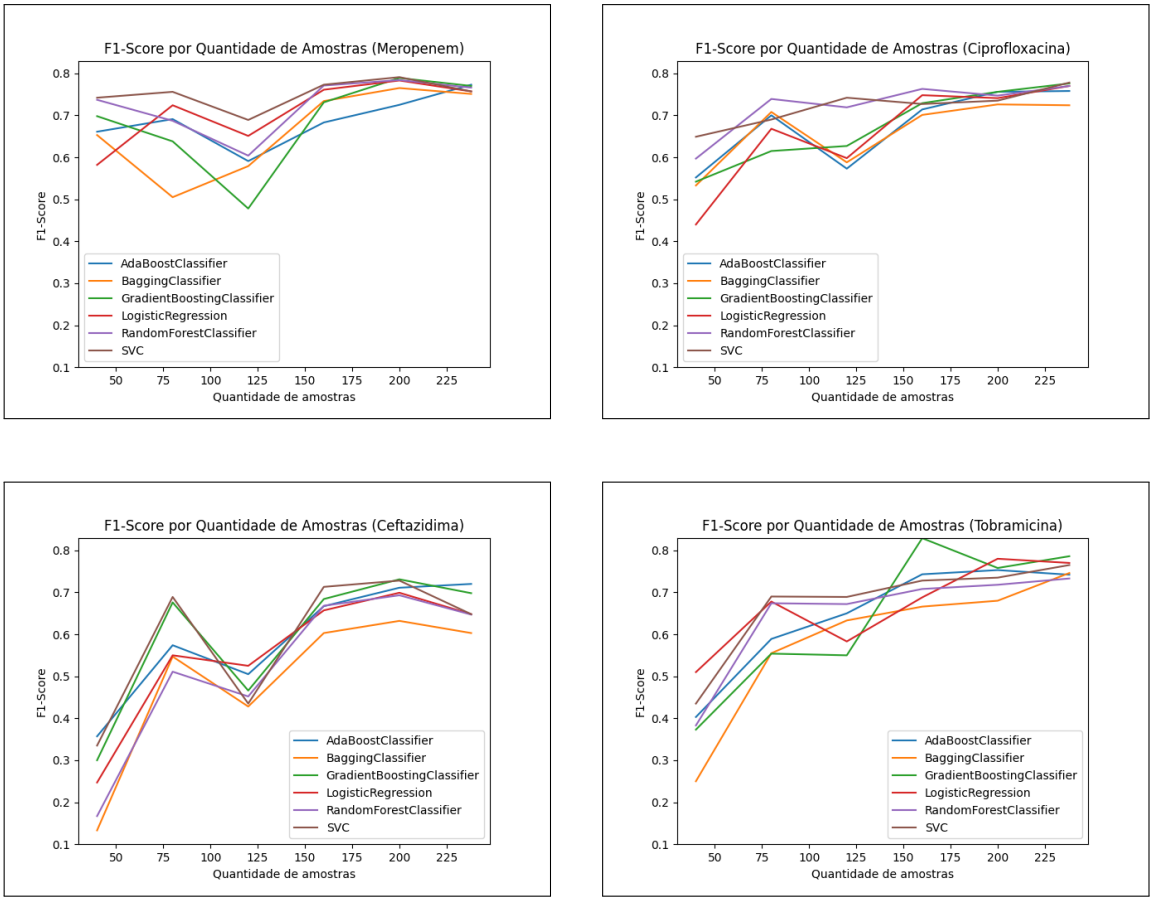
	MER	CEF	TOB	CIP
Adaboost	0.77 ±0.08	0.72 ±0.09	0.74 ±0.09	0.76 ±0.05
Bagging	0.75 ±0.04	0.60 ±0.12	0.75 ±0.13	0.72 ±0.08
GradientBoost	0.77 ±0.07	0.70 ±0.08	0.79 ±0.07	0.78 ±0.07
SVC	0.76 ±0.10	0.65 ±0.07	0.77 ±0.05	0.78 ±0.05
Regressão Logistica	0.76 ±0.10	0.65 ±0.08	0.77 ±0.06	0.77 ±0.05
Random Forest	0.77 ±0.08	0.65 ±0.08	0.73 ±0.06	0.77 ±0.04

Fonte: Autoral

3.1 ANÁLISE DO F1-SCORE POR QUANTIDADE DE AMOSTRAS PARA CADA ANTIBIÓTICO (K=9)

Com a análise dos gráficos da Figura 14 é possivel perceber o grande impacto que a quantidade de amostras tem no F1-score (k=9). Os valores foram variados de 40 a 238 com intervalos de 40 amostras.

Figura 13 – F1-score x Quantidade de Amostras.



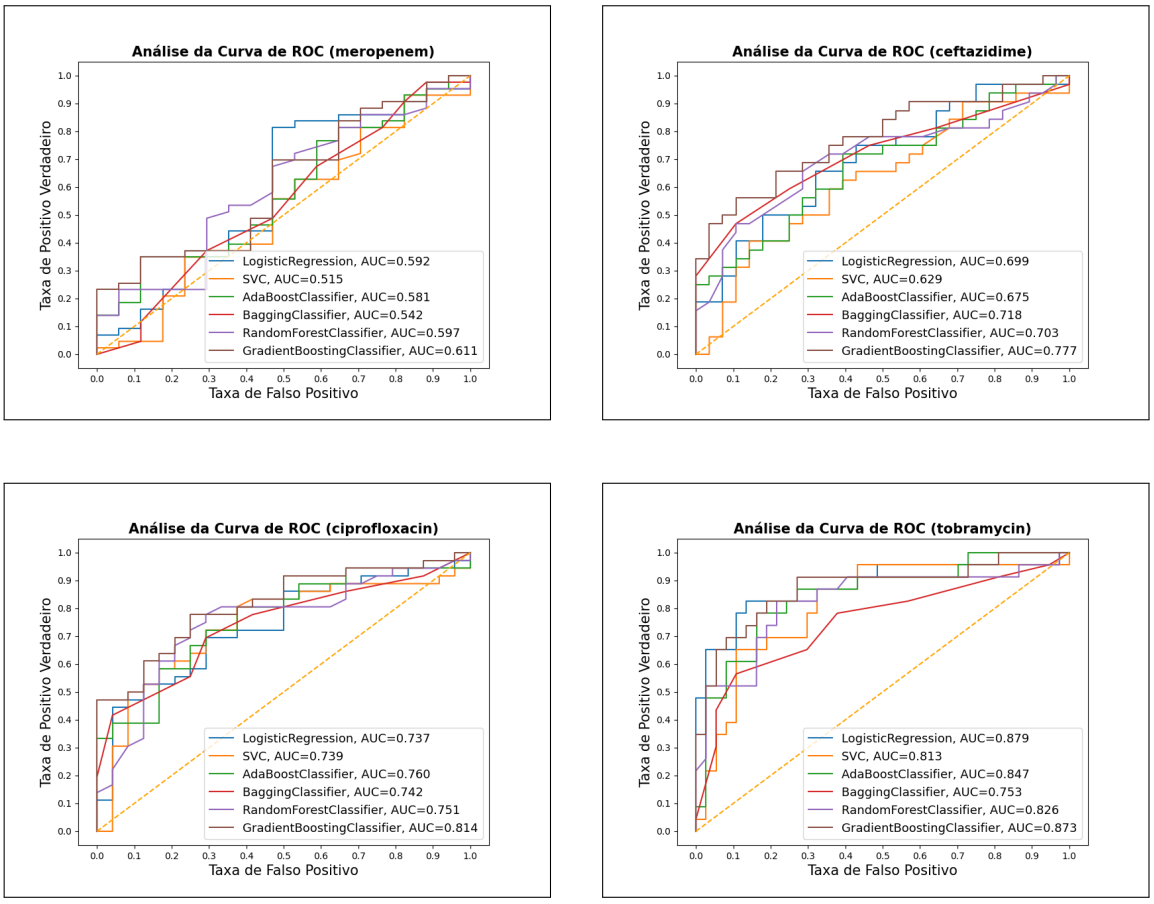
Fonte: Autoral.

O crescimento desacelerado das curvas indicam que o aumento da quantidade de amostras implica no aumento do valor do F1-Score, porém acredita-se que o limiar de estabilização das curvas, ou seja, a quantidade de amostras que fariam com que o crescimento fosse mínimo ou nulo esta próximo, haja vista a estabilização das ultimas 80 amostras. Essa ideia é provada onde foi feito testes com quantidades de amostras até 5 vezes maiores que a deste trabalho (SANTERRE *et al.*, 2016).

3.2 ANÁLISE DA CURVA ROC DOS ALGORITMOS PARA CADA ANTIBIÓTICO (K=9)

Por fim, a análise da área embaixo da curva ROC apresentou os melhores resultados para o antibiótico Tobracimina e os piores para o Meropenem, com valores entre 0.7-0.8 e entre 0.5-0.6 respectivamente.

Figura 14 – Curva ROC.



Fonte**:** Autoral.

Na tabela 4 podemos observar que o algoritmo que obteve melhor desempenho, considerando todos os antibióticos, foi classificador por aumento de gradiente (GradientBoost).

Tabela 4 – ROC-AUC por Algoritmo.

	MER	**CEF**	**TOB**	**CIP**
Adaboost	0.581	0.675	0.847	0.760
Bagging	0.542	0.718	0.753	0.742
GradientBoost	0.611	0.777	0.873	0.814
SVC	0.515	0.629	0.813	0.739
Regressão Logistica	0.592	0.699	0.879	0.737
Random Forest	0.597	0.703	0.826	0.751

Fonte**:** Autoral

4 CONCLUSÕES

Os resultados do trabalho comprovaram o grande potencial que a inteligência artificial e mais especificamente o aprendizado de máquina tem na área da predição de resistência antimicrobiana. Consequentemente, foi preenchido o vácuo que ainda permanecia nesse campo de estudo, uma vez que a predição de resistência antimicrobiana com aprendizagem de máquina utilizando exclusivamente a análise k-mer em cima do sequenciamento do genoma completo para bactérias multirresistentes como a Pseudomonas aeruginosa ainda não havia sido explorada.

4.1 CONTRIBUIÇÕES DO TRABALHO

As métricas resultantes dos treinamentos dos algoritmos apresentaram valores promissores para auxiliar no diagnóstico de infecções e para a administração personalizada de medicamentos. Os valores médio do F1-score e do AUC entre 0.7-0.8 foram considerados resultados promissores, visto que trabalhos semelhantes obtiveram resultado de metrica F1 nessa faixa (KHALEDI; WEIMANN; SCHNIEDERJANS, 2020). Acredita-se que essa diferênça se dê pelo fato de a P. aeruginosa ser não só uma bactéria multirresistente mas que também sofre intensas mutações, a influência desses aspectos nos resultados finais pode ser observada em (KOS *et al.*, 2015). Por outro lado, contornar esses impedimentos se mostrou possível com aumento do poder computacional e também do número de amostras. Além disso, a adição de outras caracteristicas genotípicas (KHALEDI; WEIMANN; SCHNIEDERJANS, 2020) e também com diferentes formas de codificação já se provaram alternativas eficazes para esse melhoramento.

4.2 LIMITAÇÕES

As limitações encontradas durante o desenvolvimento do trabalho foram primeiramente, a dificuldade na obtenção dos dados, uma vez que para conseguir dados de amostras novas seria necessário recursos materiais e mão de obra especializada durante a coleta das cepas bacterianas. Outro limitador foi a capacidade computacional, que já se demonstrou essencial no objetivo de conseguir resultados de grande confiabilidade.

4.3 TRABALHOS FUTUROS

Para trabalhos futuros, sugere-se a exploração do efeito de novas características genéticas nos treinamentos dos algorítmos de aprendizagem de máquina, como a utilização de

genes de resistência e marcadores genéticos conhecidos. Outra grande contribuição poderá ser o estudo e desenvolvimento de novas formas de codificação do sequenciamento genético para obter um melhor desempenho computacional, assim podendo produzir mais com menos recursos.

REFERÊNCIAS

ALBERTS, B.; JOHNSON, A.; LEWIS, J.; RAFF, M. **The Molecular Biology of the Cell**. Fourth edition. New York: Garland Science, 2002.

ARANGO, A. *et al.* Deeparg: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. **Microbiome**, Vol. 6, p. 1–15, 2018.

AZIZ, N.; AKHIR, E. A. P.; AZIZ, I. A. A study on gradient boosting algorithms for development of ai monitoring and prediction systems. **International Conference on Computational Intelligence**, 2020.

BREIMAN, L. Bagging predictors. **Machine Learning**, Vol. 24, p. 123–140, 1996.

BROWN, T. A. **Genomes 4**. New York and London: Garland Science, 2018.

BUERMANS, H.; DUNNEN, J. den. Next generation sequencing technology: Advances and applications. **Biochimica et Biophysica Acta**, Amsterdã, Holanda, p. 1932–1941, July 2014.

CHANG, C.-C.; LIN, C.-J. A library for support vector machines. **Department of Computer Science**, National Taiwan University, Taipei, Taiwan, 2022.

COLBY, B. Whole genome sequencing cost. **Sequencing: Outsmart Your Genes**, Disponível em: https://sequencing.com/education-center/whole-genome-sequencing/whole-genome-sequencing-cost. Acesso em: 10/12/2022, 2022.

DOERN, G. V. Antimicrobial susceptibility testing. **Journal of Clinical Microbiology**, Washington, Vol. 49, n. 9, p. S4, Sept. 2011.

FERNANDES, A. A. T.; FILHO, D. B. F.; ROCHA, E. C. da. Read this paper if you want to learn logistic regression. **Revista de Sociologia e Política**, Vol. 28, n. 74, 2020.

GUNASEKARAN, H.; RAMALAKSHMI, K.; AROKIARAJ, A. R. M. Analysis of dna sequence classification using cnn and hybrid models. **Computational and Mathematical Methods in Medicine**, Vol. 2021, 2021.

IDEXX. **Microbiology guide to interpreting minimum inhibitory concentration (MIC)**. Disponível em: https://www.idexx.com/files/microbiology-guide-interpreting-mic.pdf: [s.n.], 2019.

KHALEDI, A.; WEIMANN, A.; SCHNIEDERJANS, M. Predicting antimicrobial resistance in pseudomonas aeruginosa with machine learning-enabled molecular diagnostics. **EMBO Molecular Medicine**, Vol. 12, 2020.

KOS, V.; DéRASPE, M.; MCLAUGHLIN, R.; WHITEAKER, J.; ROY, P. The resistome of pseudomonas aeruginosa in relationship to phenotypic susceptibility. **Antimicrob Agents Chemother**, Vol. 59, p. 427 – 436, 2015.

KOUCHAKI, S. *et al.* Application of machine learning techniques to tuberculosis drug resistance analysis. **Bioinformatics**, Vol. 35, p. 2276–2282, 2019.

LV, J.; DENG, S.; ZHANG, L. A review of artificial intelligence applications for antimicrobial resistance. **Biosafety and Health**, Amsterdã, Holanda, p. 1–10, August 2020.

MAHESH, B. Machine learning algorithms - a review. **International Journal of Science and Research**, Vol. 9, 2018.

NURK, S.; KIRSCHE, M.; RHIE, A. The complete sequence of a human genome. **Science**, New York Avenue NW, Washington, p. 1–10, April 2022.

ONATE, F. P.; BATTO, J.-M.; JUSTE, C. Quality control of microbiota metagenomics by k-mer analysis. **BMC Genomics**, Vol. 16, p. 183, 2015.

O'NEILL, J. Tackling a crisis for the health and wealth of nations. **Review on Antimicrobial Resistance**, 2014.

RAY, S. A quick review of machine learning algorithms. **International Conference on Machine Learning, Big Data, Cloud and Parallel Computing**, India,, 2019.

REN, Y.; CHAKRABORTY, T.; DOIJAD, S. Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. **Bioinformatics,**, Vol. 38(2), p. 325–334, 2022.

SANTERRE, J. W.; DAVIS, J. J.; XIA, F.; STEVENS, R. Machine learning for antimicrobial resistance. **Machine Learning in Social Good Applications**, New York, NY, 2016.

STOKES, J. M. *et al.* A deep learning approach to antibiotic discovery. **Cell**, Vol. 180, p. 688–702, 2020.

TSAI, J.-K.; HUNG, C.-H. Improving adaboost classifier to predict enterprise performance after covid-19. **mathematics**, Vol. 9, p. 2215., 2021.

VELTRO, D.; ET al. Deep learning improves antimicrobial peptide recognition. **Bioinformatics**, Vol. 34, p. 2740–2747, 2018.

VENTOLA, L. The antibiotic resistance crisis: part 1: causes and threats. **P T**, v. 40, p. 277–83, 2015.

WATTAM, A.; ABRAHAM, D.; DALAY, O.; DISZ, T. Patric, the bacterial bioinformatics database and analysis resource. **Nucleic Acids Res**, Vol. 45, p. D581–D591, 2017.

WATTAM, A.; DAVIS, J.; ASSAF, R.; BOISVERT, S. Improvements topatric, the all-bacterial bioinformatics database and analysis resource center. **Nucleic Acids Res**, Vol. 45, p. D535–D542, 2017.

WHO. Global action plan on antimicrobial resistance. **World Health Organization**, 2015.

WHO. Who publishes list of bacteria for which new antibiotics are urgently needed. **World Health Organization**, Disponível em:https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed, 2017.

YANG, Y. *et al.* Machine learning for classifying tuberculosis drug-resistance from dna sequencing data. **Bioinformatics**, Vol.34, p. 1666–1671, 2018.

ZHANG. **What is FASTA format?.** Disponível em: https://zhanggroup.org/FASTA/. Acesso em: 18/11/2022: [s.n.], 2018.

ZHOU, Z.-H. Machine learning. **Springer**, Austra, Vol., p. 62–65, 2021.