UNIVERSIDADE FEDERAL DO PARÁ
NÚCLEO DE DESENVOLVIMENTO AMAZÔNICO EM ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

GISLENNE DA SILVA MOIA

PredictModelGUI: Ferramenta para classificação de genes essenciais através de técnicas de aprendizado de máquina.

Tucuruí PA

2025

GISLENNE DA SILVA MOIA

PredictModelGUI: Ferramenta para classificação de genes essenciais através de técnicas de aprendizado de máquina.

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada do Núcleo de Desenvolvimento Amazônico em Engenharia, da Universidade Federal do Pará, como requisito para a obtenção do título de Mestre em Computação Aplicada.

Orientador: Prof. Dr. Adonney Allan de Oliveira Veras
Coorientador: Prof. Dr. Cleison Daniel Silva

Tucuruí PA
2025

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

M712p Moia, Gislenne da Silva.
PredictModelGUI: Ferramenta para classificação de genes essenciais através de técnicas de aprendizado de máquina. / Gislenne da Silva Moia, . — 2025.
51 f. : il. color.

Orientador(a): Prof. Dr. Adonney Allan de Oliveira Veras
Coorientador(a): Prof. Dr. Cleison Daniel Silva
Dissertação (Mestrado) - Universidade Federal do Pará, Núcleo de Desenvolvimento Amazônico em Engenharia, Mestrado Profissional em Computação Aplicada, Tucuruí, 2025.

1. Classificação de Genes Essenciais. 2. Aprendizado de Máquina. 3. Interface Gráfica (GUI). I. Título.

CDD 005

GISLENNE DA SILVA MOIA

PredictModelGUI: Ferramenta para classificação de genes essenciais através de técnicas de aprendizado de máquina.

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada do Núcleo de Desenvolvimento Amazônico em Engenharia, da Universidade Federal do Pará, como requisito para a obtenção do título de Mestre em Computação Aplicada.

Orientador: Prof. Dr. Adonney Allan de Oliveira Veras
Coorientador: Prof. Dr. Cleison Daniel Silva

Aprovada em 06 de Junho de 2025.

BANCA EXAMINADORA:

gov.br Documento assinado digitalmente
ADONNEY ALLAN DE OLIVEIRA VERAS
Data: 25/06/2025 18:31:24-0300
Verifique em https://validar.iti.gov.br

_____
Prof. Dr. Adonney Allan de Oliveira Veras - PPCA/UFPA - Orientador

gov.br Documento assinado digitalmente
CLEISON DANIEL SILVA
Data: 28/06/2025 15:41:33-0300
Verifique em https://validar.iti.gov.br

_____
Prof. Dr. Cleison Daniel Silva - PPCA/NDAE/UFPA - Coorientador

gov.br Documento assinado digitalmente
OTAVIO NOURA TEIXEIRA
Data: 03/07/2025 09:54:08-0300
Verifique em https://validar.iti.gov.br

_____
Prof. Dr. Otávio Noura Teixeira - PPCA/NDAE/UFPA - Examinador Interno

gov.br Documento assinado digitalmente
DIEGO ASSIS DAS GRACAS
Data: 03/07/2025 11:29:04-0300
Verifique em https://validar.iti.gov.br

_____
Prof. Dr. Diego Assis das Graças - PPGBM/UFPA - Examinador Externo

Tucuruí PA
2025

PredictModelGUI: Ferramenta para classificação de genes essenciais através de técnicas de aprendizado de máquina.

Gislenne da Silva Moia[1], Victória Cardoso dos Santos[2], Saed Silva Sousa[3], Walter de Barros Gomes Netto[3], Rafael Azevedo Baraúna[4], Diego Assis das Graças[4], Artur Silva[5], Cleison Daniel Silva[1], Adonney Allan de Oliveira Veras[3,4*]

[1]Programa de Pós-Graduação em Computação Aplicada, Universidade Federal do Pará campus Tucuruí (CAMTUC-UFPA), Pará, Brasil
[2]Faculdade de Engenharia de Computação - Universidade Federal do Pará (UFPA) campus Tucuruí Pará, Brasil
[3]Faculdade de Computação - FACOMP, UFPA / CCAST.
[4]Laboratório de Engenharia Biológica, Parque de Ciência e Tecnologia, Guamá, Belém Pará, Brasil.
* Autor correspondente
E-mail: allanveras@ufpa.br (AAOV)

RESUMO

As tecnologias de sequenciamento de DNA proporcionaram avanços significativos no conhecimento sobre o conteúdo gênico de inúmeros organismos, desde microrganismos até seres humanos. Dentre as análises realizadas pelas Ciências Ômicas, a Anotação se destaca como uma das mais importantes. Conceitualmente, esse processo consiste na inferência de informações biológicas a partir de sequências genômicas, o que permite aos Pesquisadores compreender a função de produtos genéticos, como os genes — Unidades Básicas da Hereditariedade responsáveis por características físicas e hereditárias de um organismo. Alguns genes desempenham funções vitais, pois codificam proteínas ou RNAs essenciais para processos como o Metabolismo Celular, que participam em vias cruciais como a Glicólise e o Ciclo do Ácido Tricarboxílico. As Plataformas de Sequenciamento passaram a gerar grandes volumes de dados, o que impulsionou avanços nas Áreas Ômicas e fomentou o desenvolvimento de métodos computacionais voltados às mais diversas análises. Mais recentemente, técnicas de Machine Learning e Inteligência Artificial têm sido aplicadas a esses dados, com estudos que demonstram a eficácia de abordagens inspiradas na Biologia. Esses modelos não exigem programação baseada em regras, embora sua criação ainda requeira habilidades avançadas em Programação e Computação. Com o objetivo de contribuir para a solução desse desafio, este estudo apresenta o *PredictModelGUI*, uma interface gráfica desenvolvida em Python que implementa nove modelos para classificar Genes Essenciais. A

interface permite importar conjuntos de dados, re-treinar os modelos e ajustar parâmetros. As informações são armazenadas no banco de dados do *software*, o que assegura rastreabilidade e proporciona uma ferramenta simples e intuitiva para testar diferentes configurações. Disponível em: Predict Model GUI.
Palavras-chave: Classificação de Genes Essenciais, Aprendizado de Máquina, Interface Gráfica (GUI).

ABSTRACT

DNA sequencing technologies have provided significant advances in the understanding of the genetic content of numerous organisms, ranging from microorganisms to humans. Among the analyses performed in the Omics Sciences, Annotation stands out as one of the most important. Conceptually, this process consists of inferring biological information from genomic sequences, which allows researchers to understand the function of genetic products, such as Genes — the Basic Units of Heredity responsible for the physical and hereditary characteristics of an organism. Some Genes perform vital functions by encoding Proteins or RNAs essential for processes such as Cellular Metabolism, which participate in crucial pathways like Glycolysis and the Tricarboxylic Acid Cycle. Sequencing Platforms have started to generate large volumes of data, which has driven advances in the Omics fields and fostered the development of computational methods aimed at diverse analyses. More recently, Machine Learning and Artificial Intelligence techniques have been applied to these data, with studies demonstrating the effectiveness of biology-inspired approaches. These models do not require rule-based programming, although their creation still demands advanced skills in Programming and Computing. To contribute toward solving this challenge, this study presents PredictModelGUI, a graphical interface developed in Python that implements nine models to classify Essential Genes. The interface allows importing datasets, re-training models, and adjusting parameters. The information is stored in the software database, which ensures traceability and provides a simple and intuitive tool to test different configurations. Available at: Predict Model GUI.
Keywords: Essential Gene Classification. Machine Learning. GUI (Graphical User Interface).

INTRODUÇÃO

Plataformas de Sequenciamento são dispositivos utilizados para identificar as bases nucleotídicas — adenina (A), guanina (G), citosina (C) e timina (T) — em amostras de DNA. Esses dispositivos são divididos em três gerações: 1ª geração, 2ª geração (conhecida como NGS – *Next Generation Sequencing*) e 3ª geração (Satam et al., 2023; Mandlik et al., 2024).

O desenvolvimento tecnológico dessas plataformas permitiu a geração de grandes

volumes de dados. A NovaSeq 6000, da Illumina[1], por exemplo, pode produzir até 3 terabytes de dados brutos, o que corresponde a 10 bilhões de leituras. Já a PromethION2 Solo, da *Oxford Nanopore*[2], pode gerar até 580 gigabytes de dados. Entre as principais características desses dispositivos estão o baixo custo por base sequenciada, a alta velocidade na geração de informações e a elevada precisão (Hu et al., 2021; Lee, 2023; Lu et al., 2025).

A produção desse volume significativo de dados impulsionou o desenvolvimento de ferramentas computacionais para auxiliar em diversas análises, como o processo de anotação. A anotação tem ampliado o conhecimento sobre o conteúdo gênico de diferentes organismos, por meio da inferência de informações biológicas associadas aos produtos gênicos, como proteínas hipotéticas, pseudogenes, produtos com função descrita, tRNA, rRNA e, por fim, produtos com sigla de gene. O gene é compreendido como a unidade fundamental, física e funcional da hereditariedade, responsável pelas características físicas e hereditárias do fenótipo de um organismo. Os genes são compostos por nucleotídeos, considerados subunidades do DNA (ácido desoxirribonucleico) ou do RNA (ácido ribonucleico), e apresentam três componentes principais: uma pentose (açúcar de cinco carbonos), uma base nitrogenada e um ou mais grupos fosfato (Guigó, 2023).

Alguns desses genes são considerados essenciais para processos vitais, como a replicação do DNA, a transcrição e tradução de proteínas e o metabolismo celular. Genes Essenciais também desempenham um papel crucial no transporte e na absorção de nutrientes, codificação das proteínas responsáveis pela importação de nutrientes vitais e pela exportação de resíduos. Esse mecanismo é fundamental para a sobrevivência do Microrganismo em seu ambiente (Liang et al., 2024).

A análise dessas informações contribui para o enriquecimento de bancos de dados biológicos públicos, como o NCBI[3] (*National Center for Biotechnology Information*) e o DEG[4] (*Database of Essential Genes*) (Liang et al., 2024). Métodos computacionais tradicionais, como ferramentas baseadas em alinhamento por similaridade (por exemplo, *BLAST*[5]) ou regras fixas de classificação, enfrentam dificuldades para processar o grande volume e a complexidade das informações disponíveis nesses bancos. Essas abordagens geralmente não são escaláveis, exigem pré-processamento manual e apresentam limitações ao lidar com dados incompletos ou altamente variados. Além disso, métodos baseados exclusivamente em homologia não conseguem identificar genes sem similaridade conhecida, enquanto algoritmos que utilizam limiares fixos — como conteúdo de GC ou tamanho da sequência — tendem a apresentar baixa precisão. Esse cenário reforça a necessidade de técnicas mais adaptativas,

[1] https://www.illumina.com
[2] https://nanoporetech.com
[3] https://www.ncbi.nlm.nih.gov/
[4] http://origin.tubic.org/deg/public/index.php
[5] https://blast.ncbi.nlm.nih.gov/Blast.cgi

como os modelos de Inteligência Artificial - IA, que permitem maior flexibilidade e desempenho.

Técnicas de Machine Learning ou Deep Learning visam acelerar as análises e fornecer resultados com alta precisão. Essas abordagens envolvem algoritmos capazes de melhorar seu desempenho com base no treinamento, sem depender de regras de programação explícitas (Burkart & Huber, 2021).

A abordagem adotada por esses algoritmos pode ser classificada em três categorias: (i) Aprendizado Supervisionado, no qual os dados são rotulados; (ii) Aprendizado Não Supervisionado, em que os dados não são rotulados e o algoritmo identifica padrões ocultos; e (iii) Aprendizado por Reforço, em que o algoritmo interage repetidamente com o ambiente para atingir um objetivo, como vencer um jogo. Essas interações permitem encontrar soluções mais otimizadas para determinados problemas (Burkart & Huber, 2021; Naidu et al., 2023).

Uma vez treinados, esses modelos eliminam a necessidade de aplicar algoritmos baseados em regras clássicas de programação para obtenção de respostas. Basta fornecer os dados ao modelo treinado para que ele gere os resultados. No entanto, a criação desses modelos ainda requer uma etapa de programação, que pode ser desafiadora para usuários com pouca experiência em Computação (Aromolaran et al., 2021; Mukhamediev et al., 2022).

Para solucionar esse problema, é proposto o *PredictModelGUI*, uma interface gráfica desenvolvida em *Python*[6] para a classificação de Genes Essenciais com o uso de técnicas de Aprendizado de Máquina. Essa ferramenta oferece nove modelos de classificação previamente treinados que permite ao usuário classificar Genes Essenciais de forma simplificada. Além disso, a interface gráfica do *PredictModelGUI* facilita o processo de treinamento dos modelos e possibilita a personalização dos parâmetros conforme a necessidade do usuário. Todas as configurações e resultados são armazenados em um Banco de Dados, para garantir a rastreabilidade e a consistência das informações.

[6] https://www.python.org/

METODOLOGIA

Como funciona

O arquivo de entrada para o *PredictModelGUI* deve estar no formato CSV (*Comma-Separated Values*), composto por colunas que representam a frequência dos aminoácidos e, na última coluna, o nome do produto, *Product Name*, que representa a classe. Esse arquivo resulta da integração de três fontes de dados: informações de anotação e aminoácidos provenientes do banco de dados DEG e arquivos de anotação no formato *GenBank* (extensão .gb) obtidos do NCBI. Para simplificar esse processo, a própria ferramenta possui um módulo dedicado à geração automática do arquivo de entrada, o que elimina a complexidade para o usuário final.

O *PredictModelGUI* disponibiliza ao usuário três funções principais: Treinamento de Modelo, Predição e *Ensemble* — uma técnica que combina os resultados de múltiplos modelos para melhorar a precisão e robustez das classificações. Os parâmetros dessas funções podem ser ajustados diretamente na interface gráfica, o que proporciona flexibilidade, controle e permite a personalização intuitiva das configurações.

Na tarefa de treinamento de modelo, o usuário pode utilizar seu próprio conjunto de dados para treinar os modelos disponíveis no banco de dados do *software*. Nessa etapa, é necessário fornecer o conjunto de dados no formato CSV e indicar qual percentual dos dados deve servir para testar os modelos. Essa tarefa ocorre em paralelo ao treinamento para avaliar a melhoria nas métricas de avaliação à medida que o processo de aprendizagem dos modelos avança. Além disso, nesta etapa, o usuário pode selecionar quais métricas de avaliação observar: Acurácia (*Accuracy*), Precisão (*Precision*), Revocação (*Recall*), Pontuação F1 (*F1-score)*, Índice Kappa, Coeficiente de Correlação de *Matthews (Matthews Correlation Coefficient),* que servem para avaliar o desempenho dos modelos.

Para realizar a predição, o usuário pode utilizar um modelo previamente treinado (obtido na etapa anterior) no formato PKL (*Python Pickle*). Além do modelo, deve carregar o conjunto de dados no formato CSV. Para facilitar esse processo, um módulo de preparação de dados está disponível, que padroniza os dados a serem utilizados no *PredictModelGUI*, como já mencionado. Vale destacar que a ferramenta possui um modelo padrão chamado *Ensemble*, uma técnica que combina os resultados de múltiplos modelos individuais para gerar predições mais precisas e robustas. Caso o usuário opte por utilizá-lo, basta fornecer o conjunto de dados para a predição.

Na etapa de *ensemble*, o usuário pode gerar um modelo personalizado baseado em seu próprio conjunto de dados. Para treinar um novo modelo *ensemble*, deve fornecer um arquivo CSV e indicar a pasta onde os modelos previamente treinados (em formato PKL) estão armazenados. Isso torna essencial o treinamento prévio dos modelos individuais, que servem

como entrada para a construção do modelo *ensemble*. Assim como na etapa de treinamento, os parâmetros e métricas de avaliação podem ser ajustados diretamente na interface do *PredictModelGUI*, o que possibilita a personalização da configuração do modelo conforme a necessidade do usuário.

Quanto aos resultados de todas as tarefas, o usuário pode optar por recebê-los localmente, ao informar o diretório onde a pasta de resultados será gerada, ou por *e-mail*, desde que forneça o endereço diretamente na interface do *PredictModelGUI*.

Os principais resultados obtidos incluem uma avaliação gráfica dos modelos, que apresenta todas as métricas selecionadas pelo usuário. Além disso, um arquivo CSV com a lista dos Genes Essenciais identificados ao final do processo de classificação é gerado. No contexto das etapas de treinamento e *ensemble*, os modelos resultantes da análise são disponibilizados no formato PKL, o que possibilita seu reuso em etapas futuras. As etapas de funcionamento estão descritas na Figura 1.

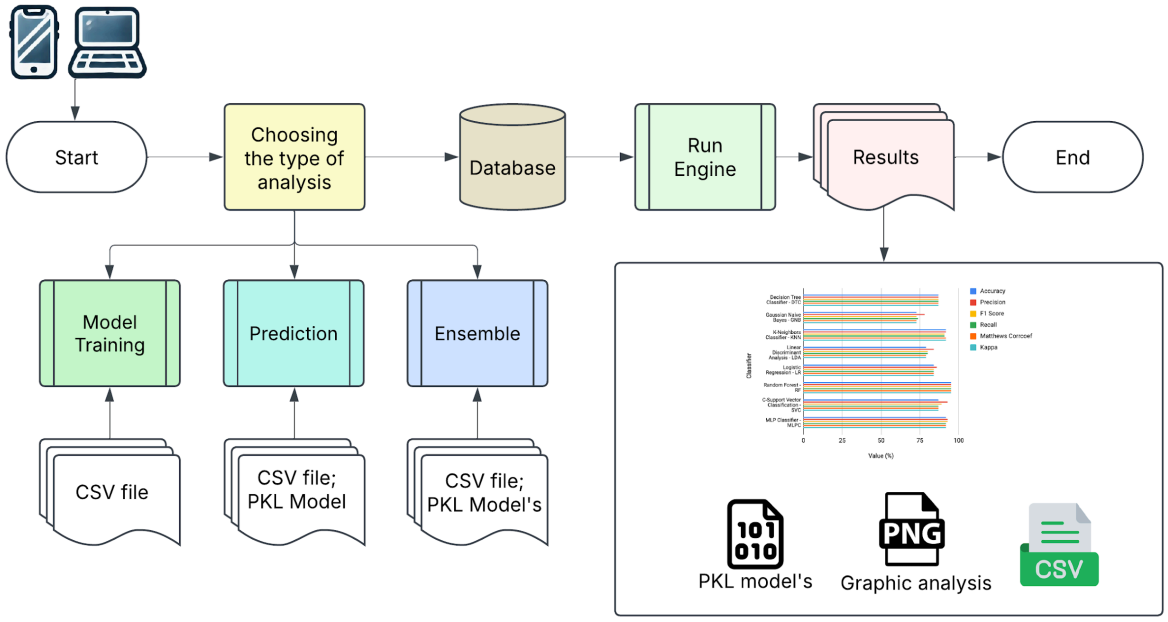

Figura 1. Pipeline do *PredictModelGUI*: A figura mostra o pipeline com as principais etapas do *software*.

Fonte de dados

Os conjuntos de dados utilizados nesta análise estão disponíveis em dois bancos de dados públicos: DEG e NCBI. As informações do DEG incluem sequências de aminoácidos e anotações de produtos gênicos bacterianos, disponíveis no formato FASTA e acompanhadas por um arquivo de anotação dessas sequências no formato CSV.

Os arquivos de anotação no formato *GenBank* foram obtidos do NCBI, que incluem 597 arquivos dos seguintes organismos: *Corynebacterium, Escherichia coli, Klebsiella, Mycobacterium, Rhodococcus, Burkholderia mallei e Salmonella*. Esses arquivos estão listados

no Anexo 01 do Material Suplementar.

Linguagem de Programação e Banco de Dados

A linguagem de programação *Python*, na versão 3.12, foi utilizada tanto para a implementação dos modelos de classificação quanto para o desenvolvimento da ferramenta *PredictModelGUI*. O *Python* é amplamente utilizado devido à sua versatilidade e à grande comunidade de desenvolvedores, especialmente nas áreas de Aprendizado de Máquina e Inteligência Artificial. O ambiente de desenvolvimento adotado foi o *PyCharm*[7].

A interface gráfica *TkintModelGUI* foi criada a partir do pacote *Tkinter*[8] na versão 8.5. O banco de dados implementado foi o *SQLite*[9], na versão 3.42.0, responsável por gerenciar as etapas de processamento e armazenar informações relacionadas aos modelos e parâmetros.

Desenvolvimento da Interface Móvel

A ferramenta móvel foi desenvolvida por meio da linguagem de programação Kotlin[10], na versão 2.0.20. Um dos principais motivos para sua escolha foi o suporte ao *Kotlin Multiplatform* (KMP), que facilita o compartilhamento de código entre diferentes plataformas. Para a criação das telas, foi adotado o *Jetpack Compose*[11], que permite construir *layouts* de forma simples e eficiente diretamente no código Kotlin.

A arquitetura MVVM (*Model-View-ViewModel*) foi selecionada para a aplicação, o que assegura uma separação clara entre a interface do usuário e a lógica da aplicação, conectadas por meio do *ViewModel*, o que promove modularidade e facilidade de manutenção. As bibliotecas utilizadas no desenvolvimento incluem: *Dokka, Ktor, Koin, Navigation Compose, Google Fonts, Kotlinx Serialization e Android Material3*.

Etapas de Pré-processamento

O pré-processamento e a criação do conjunto de dados a ser utilizado na geração dos modelos de classificação são divididos em duas etapas principais: pré-processamento do DEG e pré-processamento do NCBI.

Pré-processamento do DEG

A etapa inicial desse processo consiste no *download* dos arquivos que apresentam as sequências de aminoácidos no formato FASTA e das informações de anotação no formato CSV.

---

[7] https://www.jetbrains.com/pt-br/pycharm
[8] https://docs.python.org/pt-br/3/library/tkinter.html
[9] https://sqlite.org/
[10] https://kotlinlang.org/
[11] https://developer.android.com/compose

A Figura 2 apresenta um trecho representativo de cada um desses arquivos: (a) referente às sequências de aminoácidos (b) referente às informações de anotação. Os dados presentes nos cabeçalhos das sequências de aminoácidos são usados como referência para a construção do conjunto de dados nesta fase inicial, para assegurar a correta associação entre cada sequência de aminoácidos e sua respectiva informação de anotação.

a)

```
>DEG10010001
MENILDLWNQALAQIEKKLSKPSFETWMKSTKAHSLQGDTLTITAPNEFARDWLESRYLH
LIADTIYELTGEELSIKFVIPQNQDVEDFMPKPQVKKAVKEDTSDFPQNMLNPKYTFDTF
VIGSGNRFAHAASLAVAEAPAKAYNPLFIYGGVGLGKTHLMHAIGHYVIDHNPSAKVVYL
SSEKFTNEFINSIRDNKAVDFRNRYRNVDVLLIDDIQFLAGKEQTQEEFFHTFNTLHEES
KQIVISSDRPPKEIPTLEDRLRSRFEWGLITDITPPDLETRIAILRKKAKAEGLDIPNEV
MLYIANQIDSNIRELEGALIRVVAYSSLINKDINADLAAEALKDIIPSSKPKVITIKEIQ
RVVGQQFNIKLEDFKAKKRTKSVAFPRQIAMYLSREMTDSSLPKIGEEFGGRDHTTVIHA
HEKISKLLADDEQLQQHVKEIKEQLK
>DEG10010002
MKFTIQKDRLVESVQDVLKAVSSRTTIPILTGIKIVASDDGVSFTGSDSDISIESFIPKE
EGDKEIVTIEQPGSIVLQARFFSEIVKKLPMATVEIEVQNQYLTIIRSGKAEFNLNGLDA
DEYPHLPQIEEHHAIQIPTDLLKNLIRQTVFAVSTSETRPILTGVNWKVEQSELLCTATD
SHRLALRKAKLDIPEDRSYNVVIPGKSLTELSKILDDNQELVDIVITETQVLFKAKNVLF
FSRLLDGNYPDTTSLIPQDSKTEIIVNTKEFLQAIDRASLLAREGRNNVVKLSAKPAESI
EISSNSPEIGKVVEAIVADQIEGEELNISFSPKYMLDALKVLEGAEIRVSFTGAMRPFLI
RTPNDETIVQLILPVRTY
```

b)

DEG1001	DEG10010001	dnaA	16077069	COG0593L	DNA replication
DEG1001	DEG10010002	dnaN	16077070	COG0592L	DNA replication
DEG1001	DEG10010003	gyrB	16077074	COG0187L	DNA packaging
DEG1001	DEG10010004	gyrA	16077075	COG0188L	DNA packaging
DEG1001	DEG10010005	guaB	16077077	COG0516F	purine biosynthesis
DEG1001	DEG10010006	serS	16077081	COG0172J	tRNA synthetase
DEG1001	DEG10010007	dnaX	255767015	COG2812L	DNA replication
DEG1001	DEG10010008	tmk/tdk	255767017	COG0125F	pyrimidine biosynthesis
DEG1001	DEG10010009	holB	16077099	COG0470L	DNA replication
DEG1001	DEG10010010	metS/metG	16077106	COG0143J	tRNA synthetase
DEG1001	DEG10010011	ispE	16077114	COG1947I	isoprenoid biosynthesis
DEG1001	DEG10010012	gcaD	16077118	COG1207M	aminosugar metabolism
DEG1001	DEG10010013	prs	16077119	COG0462FE	glycolysis
DEG1001	DEG10010014	spoVC	16077121	COG0193J	RNA modification
DEG1001	DEG10010015	divIC	16077130	COG2919D	cell division

Figura 2. Pré-processamento DEG: (a) mostra um exemplo do arquivo DEG10.aa, que contém o cabeçalho de identificação da sequência seguido pelas sequências de aminoácidos. A Figura 1(b) mostra um trecho do arquivo de anotação CSV organizado em 13 colunas, composto por informações como IdDEG, IdDEGFASTA, símbolo do produto, idCOG e a anotação do produto, entre outras.

A etapa seguinte consiste na filtragem dos produtos gênicos que não possuem informação associada ao símbolo gênico, como proteínas hipotéticas ou produtos nomeados exclusivamente pelo identificador "*locus_tag*". O arquivo resultante desse processo é gerado no formato CSV, composto por um cabeçalho com as letras que representam os aminoácidos, conforme o código genético, e uma coluna chamada "*Product Name*", na qual são registradas as informações referentes ao símbolo gênico. Um exemplo desse arquivo pode ser visto na

Figura 3.

M	F	L	I	V	S	P	T	A	Y	H	Q	N	K	D	E	C	W	R	G	Product Name
3	0	5	3	7	2	2	8	2	2	2	1	1	6	2	3	0	0	5	6	rpmD
1	0	6	4	7	4	1	5	2	0	2	4	1	3	2	5	0	1	9	3	rpmD
3	1	5	5	5	2	2	6	5	0	2	1	1	5	1	4	0	0	6	5	rpmD
2	0	6	6	7	1	3	5	3	0	3	1	1	5	2	3	0	0	7	6	rpmD
3	0	5	4	6	4	1	4	1	2	3	3	3	6	1	3	1	0	4	4	rpmD
3	0	5	5	7	3	2	5	3	1	2	1	1	5	1	4	0	0	6	5	rpmD
2	0	5	6	7	1	2	4	4	0	3	3	3	6	1	4	0	0	4	6	rpmD
2	0	5	5	9	1	2	4	4	0	4	2	2	6	1	4	0	0	4	6	rpmD
1	0	6	4	9	5	0	4	5	0	1	4	4	5	2	6	0	0	7	3	rpmD
2	0	5	3	9	2	2	5	2	0	1	4	1	6	1	6	0	0	6	6	rpmD
1	0	6	5	6	8	2	5	6	0	2	4	1	4	2	3	0	1	9	6	rpmD
2	0	5	4	8	3	1	5	3	0	2	3	3	6	2	4	0	0	5	3	rpmD

Figura 3. Resultado do pré-processamento DEG: A figura mostra um trecho do conjunto de dados denominado conjunto de dados brutos DEG. Este conjunto de dados contém colunas nomeadas com os aminoácidos, onde a última corresponde ao nome do produto.

Pré-processamento do NCBI

No pré-processamento dos conjuntos de dados no formato *GenBank* (extensão .gb) obtidos do NCBI, são extraídas informações sobre os símbolos gênicos e suas respectivas sequências de aminoácidos. Para orientar essa extração, são utilizadas as *tags* "gene" (/gene) e "*translation*" (/*translation*) presentes nos arquivos, conforme mostrado na Figura 4. Em seguida, realiza-se a contagem da frequência de ocorrência dos aminoácidos em cada sequência, o que gera na geração de um arquivo chamado "*NCBI raw dataset*", estruturado no mesmo formato do conjunto de dados produzido na etapa de pré-processamento do DEG, conforme ilustrado na Figura 3.

```
CDS             43..1635
                /gene="dnaA"
                /locus_tag="CGLAU_00005"
                /inference="ab initio prediction:Prodigal:2.6"
                /inference="similar to AA sequence:UniProtKB:A0R7K1"
                /codon_start=1
                /transl_table=11
                /product="Chromosomal replication initiator protein DnaA"
                /protein_id="AQQ14009.1"
                /translation="MADQQLEALWRDLVAELLQLSERPNSPVPTLTHQQRAYLQLVKP
                VVLVDGYAILSAPHTAAKTVVEESLAPHIAAQLQARLGTPFTLAVSIAAPGVQEVGQP
                QQEPPIAQPQTEWTATQSSHTLHPPAESTYGFTAPGDQIPMGLDELAQMHARQTEESE
                RRQANAHPTVPQRIRREKPAHDPDREASLNPKYTFENFVIGSSNRFANGAAVAVAENP
                ARAYNPLFIWGGSGLGKTHLLHAAGNYARVLEPNLKIKYVSSEEFTNDYINSVRDDRQ
                ESFKRRYRNLDILMVDDIQFLEGKEGTQEEFFHTFNALHQANKQIILSSDRPPRQLTT
                LEDRLRTRFEGGLITDIQPPDLETRIAILMKKAAADGTQVSHDVLELIASQFESSIRE
                LEGALIRVSAYSSLIDEPITMDVAQVALRDILPDENDVTVTATIIKDAAAEYFQVGVE
                QLTGAGKTRHVAHARQIAMYLCRELTDLSLPKIGDEFGGKDHTTVMYADRKIRKEMTE
                NRGTYDEIQELTQLIKNRARTR"
```

Figura 4. Arquivo de anotação do NCBI *Genbank*: A figura mostra um trecho do arquivo *Genbank* no qual as *tags* "*gene*" e "*translation*" são selecionadas, que exemplifica quais informações são extraídas desses arquivos.

Os arquivos CSV gerados nas etapas de pré-processamento do DEG e do NCBI são integrados, o que resulta em um conjunto de dados enriquecido. Essa integração baseia-se nos símbolos dos Genes Essenciais identificados na etapa inicial (pré-processamento do DEG). A construção desse conjunto combinado (DEGvsNCBI) segue os seguintes critérios: (i) o símbolo gênico deve estar presente em pelo menos seis ocorrências distintas, e (ii) a frequência de aminoácidos nessas ocorrências apresenta variação em pelo menos um aminoácido, que assegura a unicidade da entrada no conjunto de dados. O produto final consiste em um arquivo CSV semelhante ao ilustrado na Figura 3, porém expandido para cobrir um maior número de observações (linhas).

Modelos de Aprendizado

Neste estudo, foram utilizados nove modelos para classificar Genes Essenciais: *Logistic Regression* (LR), *Random Forest* (RF), *Support Vector Classification* (SVC), *MLP Classifier* (MLPC), *Linear Discriminant Analysis* (LDA), *Decision Tree* (DTC), *Gaussian Naive Bayes* (NB), *K-Neighbors Classifier* (KNN), *Ensemble Learning – VotingClassifier* (EL). Cada modelo foi gerado com os valores de parâmetros descritos na Tabela 1. Ressalta-se que todos esses valores podem ser ajustados diretamente na interface gráfica do *PredictModelGUI*.

Tabela 1. Modelos e Parâmetros de Classificação: Lista de modelos, parâmetros e seus respectivos valores padrão usados na *PredictModelGUI*.

Modelos	Parâmetros
Decision Tree Classifier - DTC	*"criterion": "gini", "splitter": "best", "max_depth": None, "min_samples_split": 2, "min_samples_leaf": 1, "min_weight_fraction_leaf": 0.0, "max_features": None, "random_state": None, "max_leaf_nodes": None, "min_impurity_decrease": 0.0, "class_weight": None, "ccp_alpha": 0.0*
Gaussian Naive Bayes - GNB	*"priors": None, "var_smoothing": 1e-9*
K-Neighbors Classifier - KNN	*"n_neighbors": 5, "weights": "uniform", "algorithm": "auto", "leaf_size": 30, "p": 2, "metric": "minkowski", "metric_params": None, "n_jobs": None*
Linear Discriminant Analysis - LDA	*"solver": "svd", "shrinkage": None, "priors": None, "n_components": None, "store_covariance": False, "tol": 1e-4*

Logistic Regression - LR	*"penalty": "l2", "dual": False, "tol": 1e-4, "C": 1.0, "fit_intercept": True, "intercept_scaling": 1, "class_weight": None, "random_state": None, "solver": "lbfgs", "max_iter": 100, "multi_class": "auto", "verbose": 0, "warm_start": False, "n_jobs": None, "l1_ratio": None*
Random Forest - RF	*"n_estimators": 100, "criterion": "gini", "max_depth": None, "min_samples_split": 2, "min_samples_leaf": 1, "min_weight_fraction_leaf": 0.0, "max_features": "sqrt", "max_leaf_nodes": None, "min_impurity_decrease": 0.0, "bootstrap": True, "oob_score": False, "n_jobs": None, "random_state": None, "verbose": 0, "warm_start": False, "class_weight": None, "ccp_alpha": 0.0, "max_samples": None*
C-Support Vector Classification - SVC	*"C": 1.0, "kernel": "rbf", "degree": 3, "gamma": "scale", "coef0": 0.0, "shrinking": True, "probability": False, "tol": 1e-3, "cache_size": 200, "class_weight": None, "verbose": False, "max_iter": -1, "decision_function_shape": "ovr", "break_ties": False, "random_state": None*
MLP Classifier - MLPC	*"hidden_layer_sizes": (100,), "activation": "relu", "solver": "adam", "alpha": 0.0001, "batch_size": "auto", "learning_rate": "constant", "learning_rate_init": 0.001, "power_t": 0.5, "max_iter": 200, "shuffle": True, "random_state": None, "tol": 1e-4, "verbose": False, "warm_start": False, "momentum": 0.9, "nesterovs_momentum": True, "early_stopping": False, "validation_fraction": 0.1, "beta_1": 0.9, "beta_2": 0.999, "epsilon": 1e-8, "n_iter_no_change": 10, "max_fun": 15000*
Ensemble Learning - VotingClassifier (EL)	*estimators, *, voting='hard', weights=None, n_jobs=None, flatten_transform=True, verbose=False*

Padronização e Métricas de Avaliação

A avaliação dos modelos implementados no *PredictModelGUI* foi realizada com base nas seguintes métricas: Acurácia, Precisão, Revocação, *F1-score*, Índice Kappa, Coeficiente de Correlação de *Matthews*. Essas métricas são essenciais para fornecer uma visão abrangente do desempenho dos modelos, que ajudam a entender sua efetividade e confiabilidade, e garantem uma avaliação detalhada e precisa.

Para o processo de normalização, foi utilizado o *Standard Scaler*, que padroniza os dados para que tenham média zero e desvio padrão igual a um. Essa técnica é menos sensível a valores extremos (*outliers*) em comparação com outros métodos de escalonamento. A escolha foi feita com o objetivo de aumentar a robustez e consistência do desempenho dos modelos.

Estação de Trabalho

A estação de trabalho utilizada para gerar os modelos e realizar os testes gerais da ferramenta foi uma Penguin Computing Tempest 4400, equipada com 1024 GB de RAM, Processador AMD Opteron 6376 (64 núcleos), 4 TB de armazenamento. O sistema operacional utilizado foi o Debian 12.

RESULTADOS E DISCUSSÃO

Os resultados deste estudo estão organizados em uma análise detalhada da etapa de pré-processamento do conjunto de dados, seguida pela avaliação dos modelos preditivos por meio das métricas selecionadas. Além disso, examina-se a estratégia metodológica adotada e apresenta-se a ferramenta desenvolvida, que termina com a exibição do aplicativo móvel.

Análise da etapa de pré-processamento

As etapas de pré-processamento e os arquivos produzidos foram organizados nas seguintes fases: (1) Pré-processamento DEG, conjuntos de dados DEG10.aa e deg_annotation_p.csv: esses conjuntos de dados apresentam informações relacionadas a sequências de aminoácidos e anotações, cada um com inicialmente 26.619 observações (linhas). Após a aplicação das regras de pré-processamento, foi produzido um arquivo CSV com 12.607 observações. (2) Pré-processamento NCBI: o pré-processamento inicial do conjunto de dados NCBI gerou um arquivo CSV com 830.153 observações. (3) Cruzamento dos conjuntos de dados DEG e NCBI: esta etapa envolveu o cruzamento dos arquivos CSV DEG e NCBI, o que resultou em um arquivo com 99.818 observações.

Nesta etapa, o arquivo DEG que inclui os Genes Essenciais é usado como referência para guiar o processo, de modo que as classes (nome do produto) com a mesma frequência de aminoácidos sejam removidas. Em seguida, as classes com frequência inferior a seis representações são removidas da análise, o que gera um arquivo com 99.818 observações, equivalente a 1.720 classes distintas. A Figura 5 mostra a distribuição de frequência das classes.

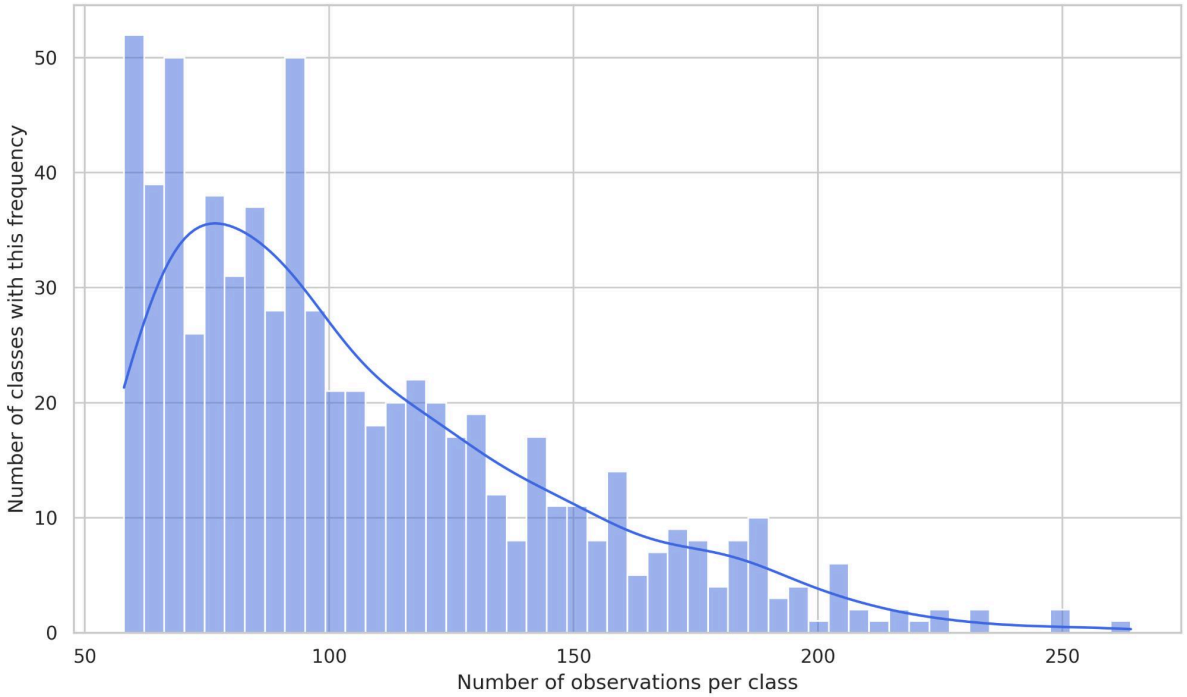

Figura 5. Distribuição da frequência de observações por classe (símbolo do gene): O eixo X representa o número de vezes que cada classe (produto com símbolo do gene na coluna "Nome do Produto") aparece no conjunto de dados. O eixo Y indica o número de classes que ocorrem com uma determinada frequência. A curva KDE (*Kernel Density Estimation*) sobreposta fornece uma estimativa suave da densidade da distribuição de frequência.

Dada a natureza desbalanceada do conjunto de dados, foi adotada uma estratégia baseada na definição da proporcionalidade das classes, que é determinada pela razão entre o número total de observações no conjunto de dados e o número de representações por classe. Como resultado dessa análise, foi estabelecido que cada classe deveria conter pelo menos 58 observações. Com base nesse critério, foi gerado um conjunto de dados composto por 74.132 observações e 696 classes, denominado *Dataset1*. As classes cuja proporcionalidade ficou abaixo desse limite foram alocadas a um segundo conjunto de dados, denominado *Dataset2*, composto por 25.686 observações e 1.024 classes.

Ao final, é realizada uma amostragem estratificada, onde se separou 1% de cada conjunto de dados para criar um conjunto de dados de validação que não participa do processo de construção do modelo. Nesse ponto, os conjuntos de dados são denominados *dataset1_toBuildModel.csv* e *dataset2_toBuildModel.csv*; para simplificar, serão denominados no manuscrito como *dataset1* e *dataset2*.

Construção de modelos e análise de métricas de avaliação

O próximo passo foi criar os modelos de classificação. Ambos os conjuntos de dados foram submetidos a uma amostragem aleatória estratificada, dividido em 80% para treinamento e

20% para teste dos modelos, seguidos pela criação dos oito modelos de classificação. Oito modelos de classificação iniciais foram desenvolvidos e, em seguida, combinados em um modelo final por meio da abordagem de Aprendizado por Conjunto. Para essa integração, foi aplicado o método *VotingClassifier*, que combina as previsões dos modelos base para melhorar a precisão da classificação.

Ao final desse processo, dezessete modelos foram criados: oito com base em cada conjunto de dados (*datasets* 1 e 2) e, por fim, o modelo resultante do agrupamento por Aprendizado de Conjunto. Os modelos foram avaliados por meio das métricas de desempenho para todos os modelos criados a partir dos dois conjuntos de dados. A análise das métricas de avaliação — Acurácia, Precisão, Recall, Pontuação F1, MCC e Kappa — é apresentada nas Figuras 6 e 7. Essas figuras mostram o desempenho dos modelos de classificação.

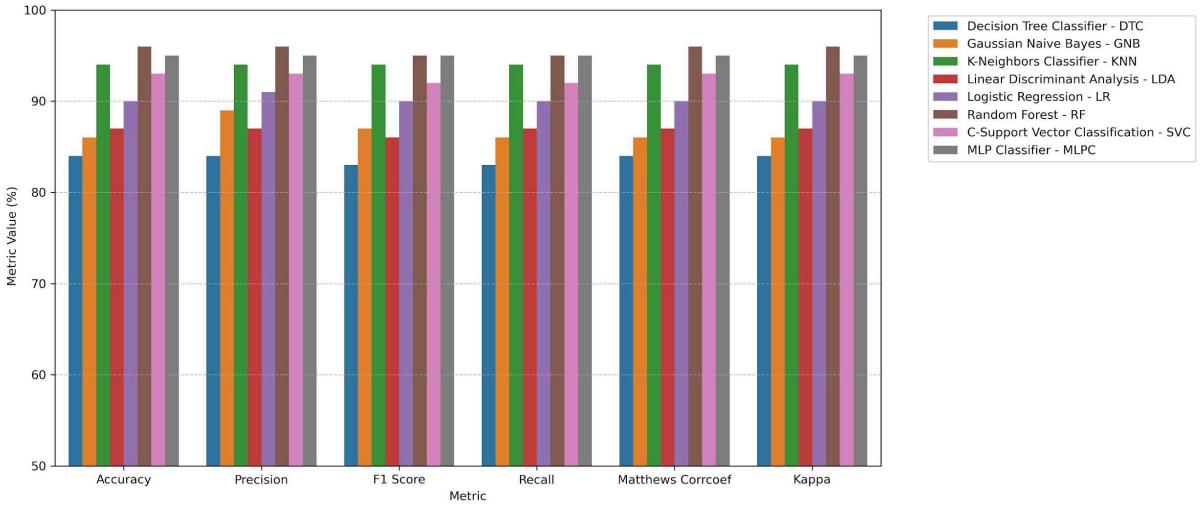

Figura 6. Avaliação dos modelos de classificação para o Conjunto de Dados 1: Análise das métricas de avaliação dos modelos criados com o Dataset1.

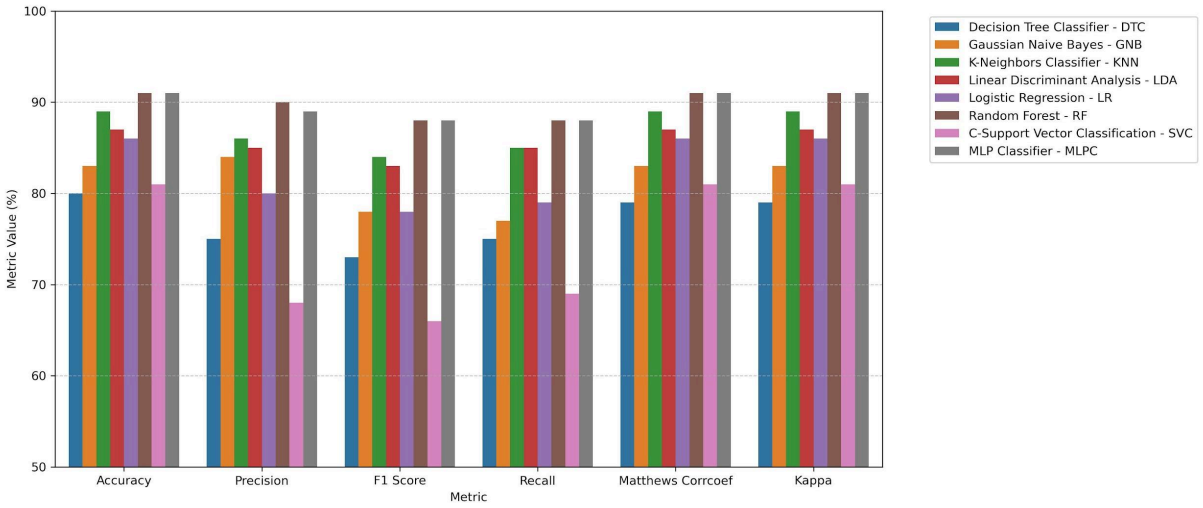

Figura 7. Avaliação dos modelos de classificação para o Conjunto de Dados 2: Análise das métricas de avaliação dos modelos criados com o Dataset2.

A análise das Figuras 6 e 7 revela um comportamento semelhante entre os modelos avaliados com diferentes conjuntos de dados. O modelo que apresentou os melhores resultados nas métricas de avaliação foi o *Random Forest*, seguido pelo *MLP Classifier,* em ambos os conjuntos. Essa consistência indica que o *Random Forest* possui desempenho superior e maior robustez nesta análise, independentemente das variações nos dados utilizados. Esse desempenho pode estar relacionado às características do próprio algoritmo, que combina múltiplas árvores de decisão e utiliza uma votação majoritária para determinar a predição final. Essa estratégia reduz o risco de sobreajuste (*overfitting*) e melhora a capacidade de generalização do modelo. Além disso, o *Random Forest* é capaz de lidar adequadamente com ruído nos dados e com o desbalanceamento entre classes, características frequentemente presentes em dados biológicos.

Embora o *MLP Classifier* também tenha apresentado desempenho elevado, sua performance tende a ser mais sensível à escolha dos hiperparâmetros e à normalização dos dados de entrada. Esses fatores podem afetar negativamente sua capacidade de generalização caso não sejam ajustados adequadamente, o que justifica sua performance ligeiramente inferior em relação ao *Random Forest.*

A boa performance de ambos os modelos, baseados em fundamentos distintos — *Random Forest* por meio de árvores de decisão e *MLP Classifier* por Redes Neurais Artificiais —, demonstra que diferentes abordagens podem ser eficazes na tarefa de classificação de Genes Essenciais. No entanto, isso não implica que a escolha do modelo seja indiferente. Cada algoritmo responde de forma diferente às variações nos dados e apresenta vantagens e limitações próprias. Dessa forma, a adoção de estratégias que exploram múltiplas técnicas, como o uso de métodos de Aprendizado de Conjunto (*ensemble*), pode contribuir para aumentar a robustez da análise, e resultar em classificações mais confiáveis e com melhor desempenho geral.

Para analisar a evolução das métricas de avaliação dos modelos de classificação criados neste estudo, foram desenvolvidos três modelos de agrupamento baseados em Aprendizagem de Conjunto. O primeiro, chamado *Ensemble Step01*, utilizou o agrupamento de modelos criados com o *dataset1*. O segundo, chamado *Ensemble Step02*, foi formado a partir de modelos baseados exclusivamente no *dataset*2. Finalmente, os dezesseis modelos foram combinados em um modelo final chamado *Ensemble Final*. É importante notar que as métricas de desempenho dos modelos de conjunto (Etapa 1, Etapa 2 e Conjunto Final) foram avaliadas com um conjunto de validação composto por uma amostra estratificada dos *dataset* 1 e 2. Essa amostra não foi usada em nenhuma etapa do processo de construção do modelo, a fim de se garantir uma avaliação imparcial e representativa do desempenho preditivo. Essa abordagem tornou possível avaliar o impacto dos diferentes agrupamentos na melhoria do desempenho do

modelo Figura 8.

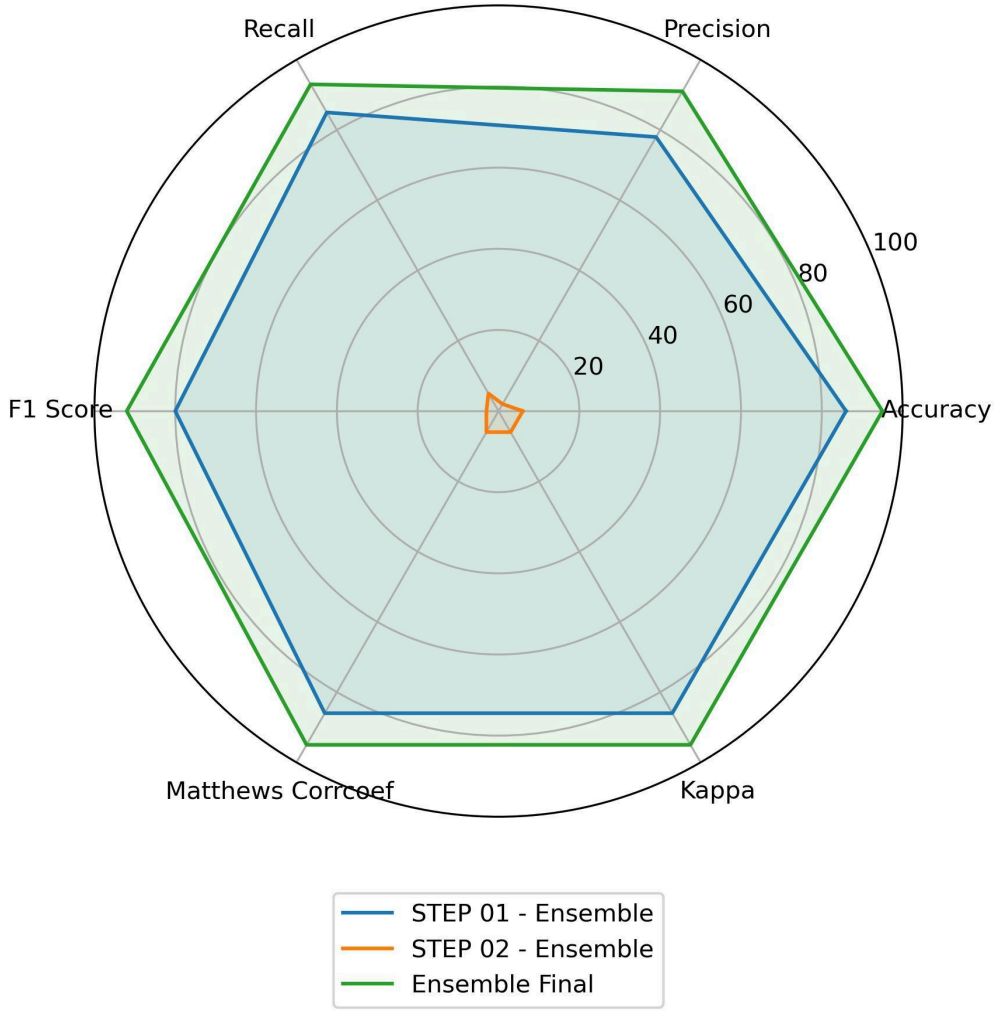

Figura 8. Avaliação do modelo de conjunto de classificação: Esta figura mostra a avaliação dos modelos gerados por meio do método de Aprendizado de Conjunto para o conjunto de oito modelos com os *dataset* 1 e 2, os quais resultaram em *ensemble_step01* e *ensemble_step02*, respectivamente.

O gráfico de radar na Figura 8 permite visualizar e comparar rapidamente o desempenho de diferentes métricas em cada estágio do conjunto. Isso facilita a análise de tendências e flutuações no desempenho do modelo em diferentes estágios de refinamento do conjunto. As métricas de avaliação que obtiveram as maiores pontuações na análise foram Precisão, MCC e Kappa, todas com 95%.

É importante enfatizar que o uso do método de aprendizagem por conjunto, que combina os modelos gerados a partir dos *dataset* 1 e 2, permite uma representação abrangente de todas as classes de Genes Essenciais do banco de dados DEG. Isso sugere que o modelo resultante é capaz de identificar um gene essencial em qualquer conjunto de dados, independentemente de sua alta ou baixa representação. Além disso, os valores das métricas de avaliação demonstram a eficácia do modelo gerado, o que reforça sua capacidade de predição precisa e confiável.

Versões para desktop e aplicativo do *PredictModelGUI*

O *PredictModelGUI* possui uma interface intuitiva e simplificada (Figura 9), desenvolvida em *Python*, com o objetivo de proporcionar uma experiência eficiente ao usuário. Ao iniciar a aplicação, uma janela introdutória é exibida, o que permite acesso imediato às funcionalidades da ferramenta. Após a inicialização, o usuário pode explorar os principais módulos de análise, conforme ilustrado na Figura 9: *Training*, que possibilita o treinamento de modelos previamente definidos e salvos no banco de dados; *Prediction*, que realiza a predição sobre um conjunto de dados e identifica automaticamente as classes (Genes Essenciais) presentes; e *Ensemble*, que possibilita a combinação de múltiplos modelos para melhorar a precisão dos resultados preditivos. Essa estrutura organizacional torna a ferramenta mais fácil de navegar e utilizar, o que promove maior eficiência e praticidade no processo de modelagem preditiva.

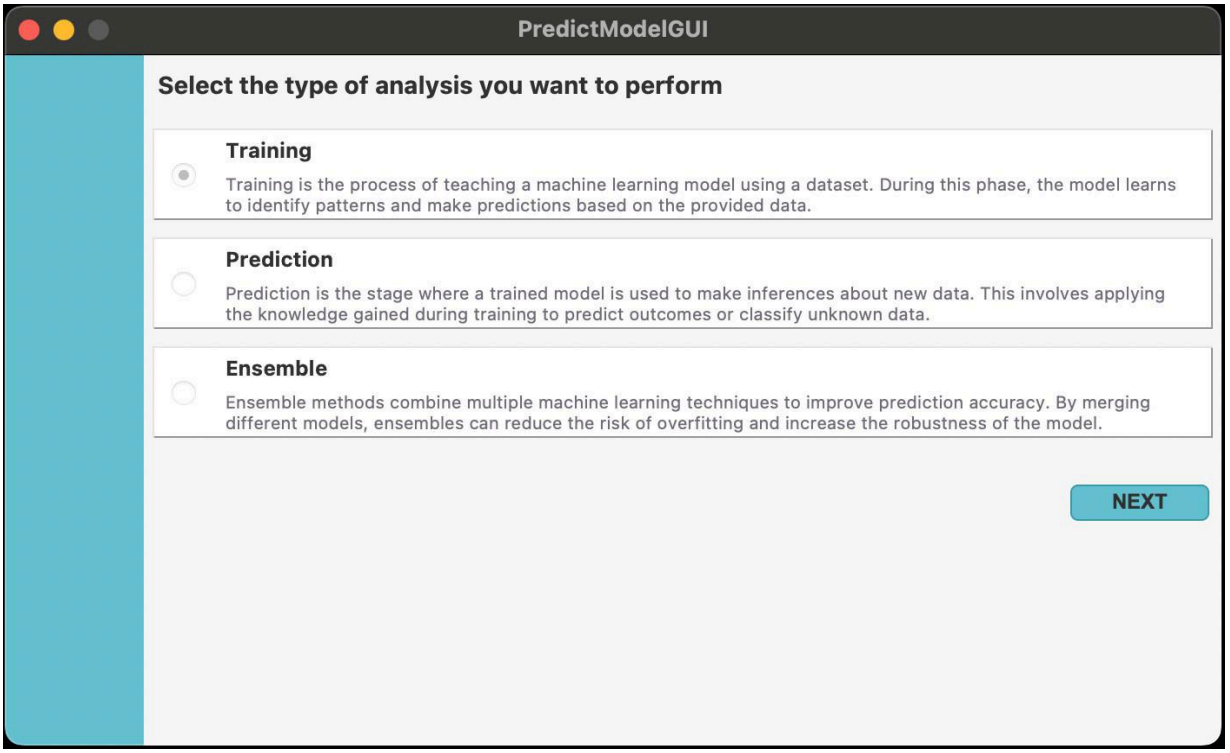

Figura 9. Janela Tipos de Análise: Esta figura exibe as opções de análise disponíveis no *PredictModelGUI*.

A seguir, são detalhados os recursos disponíveis na interface *PredictModelGUI*, com início na etapa de treinamento. Após selecionar essa opção, o usuário deve escolher entre criar um novo projeto ou carregar um existente. Caso opte por criar um novo projeto, ao inserir seu conjunto de dados, será perguntado a porcentagem do conjunto de dados a ser utilizada no processo de teste do modelo, que é de 30% por padrão na ferramenta e pode ser alterada de acordo com as necessidades do usuário. Em seguida, será exibida uma janela com todos os modelos implementados na ferramenta, conforme mostrado na Figura 10.

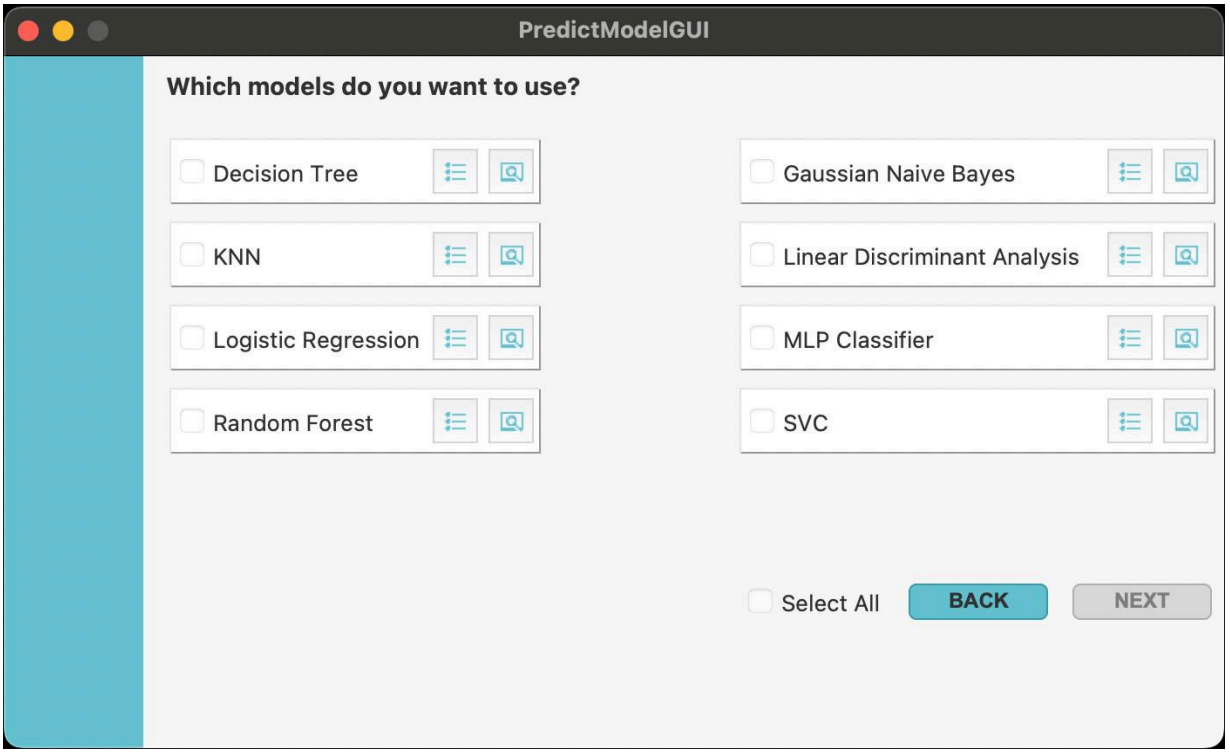

Figura 10. Lista de Modelos: A figura mostra a lista de modelos disponíveis na ferramenta *PredictModelGUI*, permite ao usuário visualizar as opções implementadas para análise.

Cabe ressaltar que todos os parâmetros podem ser modificados diretamente na interface do *software*, o que permite maior flexibilidade e personalização do processo de treinamento do modelo preditivo.

Na próxima etapa, mostrada na Figura 11, a interface exibe os parâmetros correspondentes a cada modelo selecionado na fase anterior. Nesse momento, o usuário pode optar por manter os valores padrão ou ajustá-los de acordo com as necessidades específicas de sua análise. Para facilitar essa configuração, a interface fornece acesso direto à documentação, que contém informações detalhadas sobre os parâmetros.

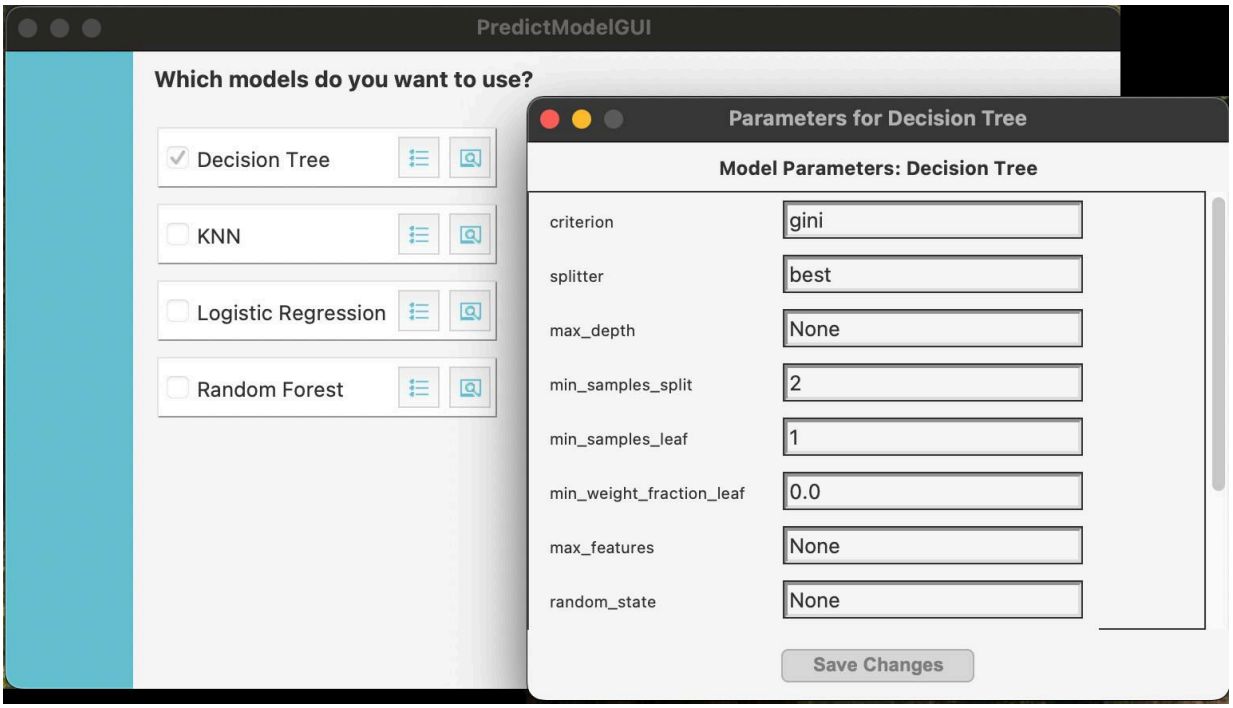

Figura 11. Lista de Parâmetros: Janela para edição de parâmetros, no exemplo são mostrados os parâmetros do modelo *Decision Tree.*

Na próxima etapa, o usuário pode selecionar as métricas que serão utilizadas para avaliar os modelos treinados. A Figura 12 mostra todas as métricas implementadas na *PredictModelGUI*, o que possibilita que o usuário defina quais serão utilizadas no processo de avaliação. Dessa forma, é possível escolher um subconjunto específico de métricas ou utilizar todas elas, a partir dos critérios adotados.

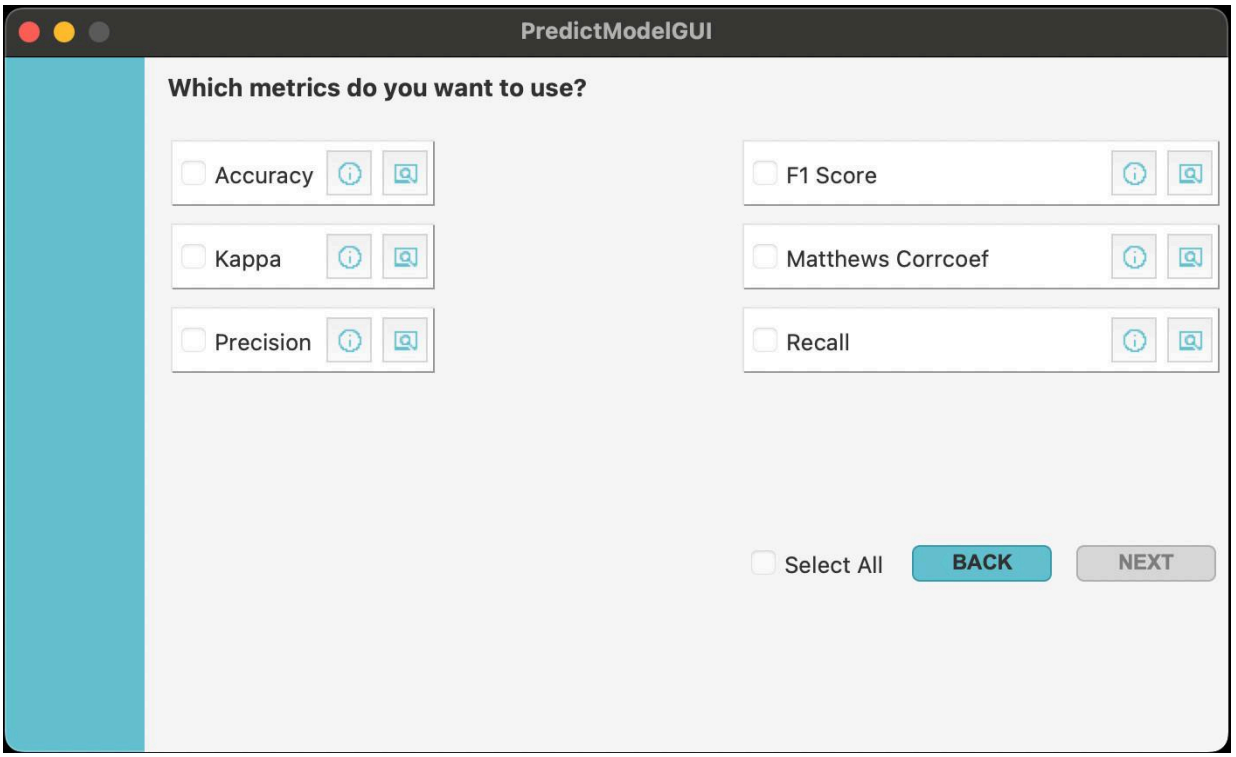

Figura 12. Lista de métricas de avaliação do modelo: Esta janela exibe todas as métricas de avaliação

disponíveis na ferramenta *PredictModelGUI*, permite ao usuário escolher as opções mais adequadas para analisar o desempenho do modelo.

Por fim, o usuário pode especificar o método de recebimento dos resultados. Se optar pelo armazenamento local, deverá indicar o diretório onde será gerada a pasta com os resultados. Alternativamente, o usuário pode optar por enviar os resultados por *e-mail*, ao informar o endereço de e-mail diretamente na interface do *software*.

Na etapa de análise, chamada Predição, o usuário pode fazer previsões a partir de um conjunto de dados em formato CSV como entrada. O *PredictModelGUI* possui um modelo padrão chamado *Ensemble*, que é compactado na primeira utilização. Portanto, antes da utilização, o *software* descompacta automaticamente o modelo e, em seguida, carrega-o na ferramenta. Esse modelo é então armazenado localmente em uma pasta específica, criada automaticamente durante o processo de descompactação.

Além disso, o equipamento utilizado deve atender aos requisitos mínimos de *hardware* para garantir o funcionamento adequado do modelo. Para facilitar essa verificação, o *PredictModelGUI* realiza automaticamente um teste de compatibilidade e informa o usuário sobre a viabilidade de utilização do modelo no dispositivo em questão.

O usuário tem a opção de utilizar um modelo previamente treinado, por meio dos procedimentos descritos na etapa de treinamento. Para isso, é necessário carregar o modelo em formato PKL, juntamente com o conjunto de dados em formato CSV. As etapas subsequentes, são compostas por edição de parâmetros, definição de métricas de avaliação e escolha do método de recebimento dos resultados, e seguem o mesmo fluxo descrito acima.

Para facilitar a padronização do arquivo de entrada na ferramenta *PredictModelGUI*, um módulo específico é fornecido para a conversão e preparação do conjunto de dados em formato CSV a partir de arquivos de anotação no formato *GenBank*. Este módulo visa garantir que os dados estejam em conformidade com os requisitos da ferramenta.

Na última etapa, correspondente à tarefa *Ensemble*, o usuário tem a opção de recriar o modelo de conjunto padrão da ferramenta através do seu conjunto de dados. Para isso, é necessário primeiro treinar os modelos individualmente, com os procedimentos descritos na etapa de treinamento. Após a conclusão desta etapa, os modelos treinados devem ser organizados no formato PKL e armazenados em um único diretório.

Na fase de *ensemble*, o usuário deve fornecer o conjunto de dados em formato CSV e indicar o diretório onde os modelos previamente treinados estão armazenados. Com base nessas informações, o modelo de conjunto será treinado.

As janelas já descritas para parametrização e definição de métricas de avaliação e resultados são as mesmas para esta etapa. É importante observar que, uma vez treinado, o modelo *ensemble* deve ser carregado sempre que o usuário desejar utilizá-lo para fazer

previsões.

Para otimizar o uso da ferramenta *PredictModelGUI*, foi desenvolvida uma versão mobile, mostrada na Figura 13, disponível na *Google Play Store*. Essa versão permite que o usuário acesse as funcionalidades do *PredictModelGUI* por meio da comunicação com uma versão *Server*. Dessa forma, foi implementada uma arquitetura cliente/servidor. Após a conclusão do processo selecionado pelo usuário, uma notificação de encerramento é enviada para o endereço de *e-mail* cadastrado.

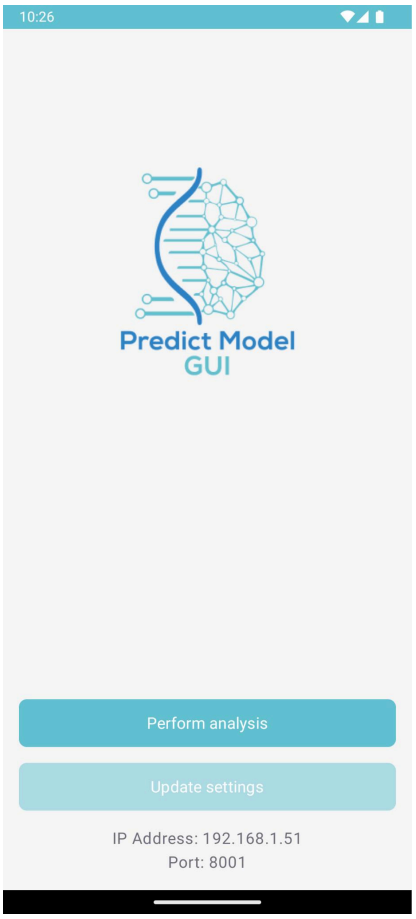

Figura 13. Versão móvel do *PredictModelGUI*: Janela principal do aplicativo *PredictModelGUI*.

O padrão visual das janelas foi mantido consistente entre as versões *desktop* e *mobile*, o que oferece uma experiência de usuário homogênea e intuitiva em ambas as plataformas.

A análise dos resultados demonstra três principais contribuições científicas do desenvolvimento da *PredictModelGUI*. Primeiramente, a metodologia inovadora adotada para a geração e o processamento do conjunto de dados melhorou a qualidade das informações, o que gerou em modelos preditivos mais precisos e confiáveis. Em segundo lugar, a implementação de uma abordagem diferenciada para a construção dos modelos de classificação permitiu uma cobertura abrangente de todas as classes genéticas essenciais, para promover melhorias significativas nas métricas de avaliação. Por fim, o desenvolvimento da ferramenta computacional permitiu a integração dos modelos em uma interface intuitiva e amigável, para reduzir a complexidade operacional para o usuário final.

Além disso, o *PredictModelGUI* possui uma versão compatível com arquiteturas cliente-servidor, o que permite a comunicação via dispositivos móveis e aumenta sua acessibilidade e aplicabilidade em diferentes contextos. Informações detalhadas sobre como executar cada etapa descrita neste manuscrito são fornecidas no manual do usuário do *PredictModelGUI*.

Por fim, é importante destacar que o desenvolvimento da ferramenta *PredictModelGUI* proporcionou uma maneira eficiente de abstrair a complexidade envolvida na construção, no treinamento e no uso de modelos de Aprendizado de Máquina. Além disso, a ferramenta amplia o acesso a essa tecnologia, especialmente para usuários sem conhecimento aprofundado de computação, e torna a aplicação mais acessível e intuitiva.

CONCLUSÃO

Neste estudo, investigou-se o uso de técnicas de Aprendizado de Máquina na classificação de Genes Essenciais. Para tanto, adotou-se uma abordagem inovadora na criação do conjunto de dados que utiliza informações de dois bancos de dados: DEG e NCBI. As informações sobre Genes Essenciais são registradas no DEG e comparadas com os dados de anotação do NCBI.

Vale ressaltar que o conjunto de dados criado exibe, ao final, a frequência de aminoácidos associada a cada produto, com um símbolo de gene que corresponde às observações do conjunto de dados. Ao criar o conjunto de dados, foi estabelecida uma estratégia para determinar o valor mínimo de proporcionalidade de representação que cada classe deveria ter. Assim, foram criados dois conjuntos de dados: um com valores iguais ou superiores ao estabelecido pela estratégia e outro com valores menores.

A próxima etapa envolveu o treinamento dos modelos com base nos conjuntos de dados criados anteriormente. No entanto, cada modelo gerado a partir de um determinado conjunto de dados contém apenas uma fração do conhecimento sobre Genes Essenciais, influenciado pela proporcionalidade das classes utilizadas para construir esses conjuntos de dados.

Essa limitação fica clara ao analisar os resultados das métricas de avaliação. A comparação dos valores obtidos para os dois conjuntos de dados mostra uma disparidade significativa entre as métricas. A título de comparação, os resultados para os *datasets* 1 e 2 foram, respectivamente: Acurácia de 86% e 6%; Precisão de 78% e 2%; F1-Score de 85% e 5%; Recall de 80% e 3%; as métricas Coeficiente de Correlação de *Matthews* (MCC) e Kappa obtiveram os mesmos valores de 86% e 6%.

Para mitigar essa discrepância, os dois modelos gerados anteriormente a partir dos conjuntos de dados foram combinados para formar um modelo final, denominado *ensemble* final. Essa técnica, amplamente utilizada em Inteligência Artificial e Aprendizado de Máquina, permite a agregação de diferentes modelos, o que proporciona um melhor desempenho

preditivo do que modelos individuais.

Essa abordagem se mostrou eficaz, como demonstrado pelas métricas de desempenho: Acurácia, MCC e Kappa atingiram 95%; Precisão atingiu 91%; F1-Score atingiu 93%; e Recall atingiu 92%. É importante observar que esses valores foram obtidos por meio do conjunto de dados de validação, não a parte de teste utilizada durante o treinamento do modelo. Esses resultados demonstram a eficiência do modelo na identificação de Genes Essenciais e reforçam a aplicação inovadora do método de Aprendizado por Conjunto de dados biológicos.

A análise e a avaliação dos modelos implementados no *PredictModelGUI* demonstraram alta eficiência na identificação de Genes Essenciais. Essa funcionalidade é aprimorada por uma interface gráfica intuitiva, que simplifica o processo de treinamento e utilização dos modelos, o que elimina a necessidade de conhecimentos avançados de computação por parte do usuário. Além disso, foi desenvolvida uma versão *mobile* da ferramenta, que envia os dados para processamento em um dispositivo remoto e, em seguida, recebe os resultados diretamente no *e-mail*, o que possibilita o uso da ferramenta de forma mais acessível, dinâmica e versátil.

É importante observar que, à medida que o usuário expande o conjunto de dados e ajusta os valores dos parâmetros, os modelos podem apresentar desempenho aprimorado, refletido em melhores métricas de avaliação. Isso indica um aumento em sua eficácia preditiva. O principal objetivo deste trabalho foi desenvolver uma ferramenta que simplifique a complexidade inerente ao desenvolvimento dos *scripts* necessários para a criação dos modelos e facilite tanto o processo de treinamento quanto a sua utilização.

AGRADECIMENTOS

Agradecimentos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Universidade Federal do Pará. Este trabalho foi financiado pela PROPESP/PAPQ. Este trabalho faz parte da pesquisa realizada pelo BIOD (Grupo de Pesquisa em Bioinformática, Ômicas e Desenvolvimento) e pelo EngBio (Laboratório de Engenharia Biológica).

CONFLITOS DE INTERESSE

Os autores declaram que nenhum conflito de interesse pode ser percebido como prejudicial à imparcialidade da pesquisa relatada.

CONTRIBUIÇÕES DOS AUTORES

GSM desenvolvimento de *software* e aplicativo móvel, redação do manuscrito, redação – revisão e edição, análise formal e redação de manuais técnicos. VCS desenvolvimento de *software*, redação de manuais técnicos e testes de *software*. SSS desenvolvimento de aplicativo

móvel. WBGN teste de *software* e aplicativo móvel. RAB, DAG e AS redação – revisão e edição. CDS redação – revisão e edição e análise formal. AAOV investigação, metodologia, papéis/redação – rascunho original e desenvolvimento de *software*. Todos os autores leram e aprovaram a versão final.

REFERÊNCIAS

Burkart, N. and Huber, M.F. (2021) 'A survey on the explainability of supervised machine learning', Journal of Artificial Intelligence Research, 70, pp. 245–317. doi:10.1613/jair.1.12228.

Guigó, R. (2023) 'Genome annotation: From human genetics to biodiversity genomics', Cell Genomics, 3(8), p. 100375. doi:10.1016/j.xgen.2023.100375.

Lee JY. The Principles and Applications of High-Throughput Sequencing Technologies. Dev Reprod. 2023 Apr;27(1):9-24. doi: 10.12717/DR.2023.27.1.9. Epub 2023 Mar 31. PMID: 38075439; PMCID: PMC10703097.

Liang, Y. et al. (2024) 'Recent advances in the characterization of essential genes and development of a database of essential genes', iMeta, 3(1). doi:10.1002/imt2.157.

Lu, Y.; Li, M.; Gao, Z.; Ma, H.; Chong, Y.; Hong, J.; Wu, J.; Wu, D.; Xi, D.; Deng, W. Advances in Whole Genome Sequencing: Methods, Tools, and Applications in Population Genomics. Int. J. Mol. Sci. 2025, 26, 372. https://doi.org/10.3390/ijms26010372.

Mandlik, J.S., Patil, A.S. and Singh, S. (2024) 'Next-generation sequencing (NGS): Platforms and applications', Journal of Pharmacy and Bioallied Sciences, 16(Suppl 1). doi:10.4103/jpbs.jpbs_838_23.

Mukhamediev RI, Popova Y, Kuchin Y, Zaitseva E, Kalimoldayev A, Symagulov A, Levashenko V, Abdoldina F, Gopejenko V, Yakunin K, et al. Review of Artificial Intelligence and Machine Learning Technologies: Classification, Restrictions, Opportunities and Challenges. Mathematics. 2022; 10(15):2552. https://doi.org/10.3390/math10152552

Naidu, G., Zuva, T., Sibanda, E.M. (2023). A Review of Evaluation Metrics in Machine Learning Algorithms. In: Silhavy, R., Silhavy, P. (eds) Artificial Intelligence Application in Networks and Systems. CSOC 2023. Lecture Notes in Networks and Systems, vol 724. Springer, Cham. https://doi.org/10.1007/978-3-031-35314-7_2

Olufemi Aromolaran, Damilare Aromolaran, Itunuoluwa Isewon, Jelili Oyelade, Machine learning approach to gene essentiality prediction: a review, Briefings in Bioinformatics, Volume 22, Issue 5, September 2021, bbab128, https://doi.org/10.1093/bib/bbab128

Satam H, Joshi K, Mangrolia U, Waghoo S, Zaidi G, Rawool S, Thakare RP, Banday S, Mishra AK, Das G, et al. Next-Generation Sequencing Technology: Current Trends and Advancements. Biology. 2023; 12(7):997. https://doi.org/10.3390/biology12070997

Taishan Hu, Nilesh Chitnis, Dimitri Monos, Anh Dinh, Next-generation sequencing technologies: An overview, Human Immunology, Volume 82, Issue 11, 2021, Pages 801-811, ISSN 0198-8859, https://doi.org/10.1016/j.humimm.2021.02.012.

MATERIAL SUPLEMENTAR

Anexo 01. Lista de organismos usados para construir os conjuntos de dados: A coluna 1 lista os nomes dos organismos e a coluna 2 lista o número de acesso.

Nome do organismo	Número de acesso
Corynebacterium Ammoniagenes Dsm 20306 9.6	CP009244
Corynebacterium Ammoniagenes Kccm 40472	CP019705
Corynebacterium Amycolatum Fdaargos 1107	CP068169
Corynebacterium Amycolatum Fdaargos 1108	CP068168
Corynebacterium Anserum 23h37-10	CP046883
Corynebacterium Aquatimens Cnctc	CP068278
Corynebacterium Argentoratense Dsm 44202	CP006365
Corynebacterium Aurimucosum Atcc 700975	CP001601
Corynebacterium Aurimucosum Fdaargos 1109	CP068166
Corynebacterium Aurimucosum Fdaargos 1110	CP068165
Corynebacterium Bovis Fdaargos 1052	CP066067
Corynebacterium Callunae Dsm 20147	CP004354
Corynebacterium Camporealensis Cip 105508	CP027001
Corynebacterium Camporealensis Dsm 44610	CP011311
Corynebacterium Casei Lmg S-19264	CP004350
Corynebacterium Coyleae Fdaargos 1425	CP077302
Corynebacterium Coyleae Fdaargos 1492	CP083648
Corynebacterium Diphtheriae 241	CP003207

Corynebacterium Diphtheriae 31a	CP003206
Corynebacterium Diphtheriae B-D-16-78	CP018331
Corynebacterium Diphtheriae Bh8	CP003209
Corynebacterium Diphtheriae Bq11	CP029644
Corynebacterium Diphtheriae C7(Beta)	CP003210
Corynebacterium Diphtheriae Cdce 8392	CP003211
Corynebacterium Diphtheriae Cn2000	CP039522
Corynebacterium Diphtheriae Dongyang	CP074413
Corynebacterium Diphtheriae Fdaargos 1552	CP085960
Corynebacterium Diphtheriae Fdaargos 1553	CP085959
Corynebacterium Diphtheriae Fdaargos 1592	CP085994
Corynebacterium Diphtheriae Fdaargos 197	CP020410
Corynebacterium Diphtheriae Hc01	CP003212
Corynebacterium Diphtheriae Th1526	CP038504
Corynebacterium Doosanense Cau 212 Dsm 45436	CP006764
Corynebacterium Falsenii Fdaargos 1493	CP083647
Corynebacterium Falsenii Fdaargos 1494	CP083646
Corynebacterium Flavescens Oj8	CP009246
Corynebacterium Freneyi Fdaargos 1426	CP077258
Corynebacterium Glaucum Dsm 30827	CP019688
Corynebacterium Glucuronolyticum Fdaargos 1053	CP066007
Corynebacterium Glucuronolyticum Fdaargos 1111	CP068162
Corynebacterium Glucuronolyticum Fdaargos 1190	CP069485
Corynebacterium Glutamicum Ar1	CP007724
Corynebacterium Glutamicum Atcc 13032	CP025533
Corynebacterium Glutamicum Atcc 13869	CP016335
Corynebacterium Glutamicum Atcc 14067	CP022614

Corynebacterium Glutamicum Atcc 21573	CP068290
Corynebacterium Glutamicum Atcc 21831	CP007722
Corynebacterium Glutamicum B253	CP010451
Corynebacterium Glutamicum B414	CP012297
Corynebacterium Glutamicum Bca	CP059382
Corynebacterium Glutamicum Be	CP053188
Corynebacterium Glutamicum C1	CP017995
Corynebacterium Glutamicum Cgmcc1.15647	CP073911
Corynebacterium Glutamicum Cicc10064	CP012298
Corynebacterium Glutamicum Cp	CP012194
Corynebacterium Glutamicum Cr101	CP080542
Corynebacterium Glutamicum Ha	CP025534
Corynebacterium Glutamicum Jh41	CP041729
Corynebacterium Glutamicum Mb001	CP005959
Corynebacterium Glutamicum Scgg1	CP004047
Corynebacterium Glutamicum Scgg2	CP004048
Corynebacterium Glutamicum Tccc11822	CP020033
Corynebacterium Glutamicum Tq2223	CP020658
Corynebacterium Glutamicum Usda-Ars-Usmarc-56828	CP013991
Corynebacterium Glutamicum Wm001	CP022394
Corynebacterium Glutamicum Xv	CP018175
Corynebacterium Glutamicum Yi	CP014984
Corynebacterium Glutamicum Zl-6	CP004062
Corynebacterium Glyciniphilum Aj 3170	CP006842
Corynebacterium Halotolerans Yim 70093 Dsm 44683	CP003697
Corynebacterium Humireducens Nbrc 106098 Dsm 45392	CP005286
Corynebacterium Imitans Dsm 44264	CP009211

Corynebacterium Jeikeium Fdaargos 328	CP022054
Corynebacterium Jeikeium Fdaargos 574	CP033784
Corynebacterium Kefirresidentii Fdaargos 1055	CP067012
Corynebacterium Kroppenstedtii Dsm 44385	CP001620
Corynebacterium Kroppenstedtii Fdaargos 1192	CP069509
Corynebacterium Kutscheri Dsm 20755	CP011312
Corynebacterium Lactis Rw2-5	CP006841
Corynebacterium Liangguodongii 2184	CP026948
Corynebacterium Lizhenjunii Zj-599	CP064954
Corynebacterium Lujinxingii Zg-917	CP061032
Corynebacterium Macginleyi 160811	CP068292
Corynebacterium Macginleyi 180208	CP068291
Corynebacterium Macginleyi Fdaargos 1195	CP069516
Corynebacterium Marinum Dsm 44953	CP007790
Corynebacterium Matruchotii Atcc 14266	CP050134
Corynebacterium Minutissimum Fdaargos 1196	CP069533
Corynebacterium Minutissimum Fdaargos 894	CP065689
Corynebacterium Minutissimum Fdaargos 992	CP065964
Corynebacterium Mustelae Dsm 45274	CP011542
Corynebacterium Nuruki S6-4	CP042429
Corynebacterium Pelargi 136-3	CP035299
Corynebacterium Phocae M408-89-1	CP009249
Corynebacterium Propinquum Fdaargos 1112	CP068161
Corynebacterium Pseudotuberculosis 39	CP015188
Corynebacterium Pseudotuberculosis Atcc 19410	CP021251
Corynebacterium Pseudotuberculosis Mb66	CP013263
Corynebacterium Striatum 216	CP024932

Corynebacterium Striatum Kc-Na-01	CP021252
Escherichia Coli Isolate Eccnb12-2	NZ_CP033635
Escherichia Coli Strain 06-00048	NZ_CP015229
Escherichia Coli Strain 06-3462	NZ_CP034794
Escherichia Coli Strain 08-3914	NZ_CP034808
Escherichia Coli Strain 105	NZ_CP028700
Escherichia Coli Strain 106	NZ_CP028698
Escherichia Coli Strain 107	NZ_CP028695
Escherichia Coli Strain 108	NZ_CP028693
Escherichia Coli Strain 109	NZ_CP028690
Escherichia Coli Strain 110	NZ_CP028687
Escherichia Coli Strain 111	NZ_CP028685
Escherichia Coli Strain 112	NZ_CP028683
Escherichia Coli Strain 113	NZ_CP028680
Escherichia Coli Strain 114	NZ_CP028677
Escherichia Coli Strain 116	NZ_CP028671
Escherichia Coli Strain 117	NZ_CP028668
Escherichia Coli Strain 118	NZ_CP028665
Escherichia Coli Strain 119	NZ_CP028662
Escherichia Coli Strain 120	CP028659
Escherichia Coli Strain 121	NZ_CP028656
Escherichia Coli Strain 122	NZ_CP028654
Escherichia Coli Strain 123	NZ_CP028652
Escherichia Coli Strain 124	NZ_CP028650
Escherichia Coli Strain 130	NZ_CP028647
Escherichia Coli Strain 131	NZ_CP028644
Escherichia Coli Strain 132	NZ_CP028641

Escherichia Coli Strain 133	NZ_CP028638
Escherichia Coli Strain 134	NZ_CP028635
Escherichia Coli Strain 135	NZ_CP028632
Escherichia Coli Strain 136	NZ_CP028629
Escherichia Coli Strain 137	NZ_CP028626
Escherichia Coli Strain 138	NZ_CP028623
Escherichia Coli Strain 139	NZ_CP028620
Escherichia Coli Strain 140	NZ_CP028617
Escherichia Coli Strain 141	NZ_CP028614
Escherichia Coli Strain 143	NZ_CP028607
Escherichia Coli Strain 144	NZ_CP028603
Escherichia Coli Strain 150	NZ_CP028592
Escherichia Coli Strain 155	NZ_CP018237
Escherichia Coli Strain 2009c-3554	CP034803
Escherichia Coli Strain 2009c-4687	NZ_CP034799
Escherichia Coli Strain 2010c-3142	NZ_CP034801
Escherichia Coli Strain 2010c-3347	NZ_CP034806
Escherichia Coli Strain 2011c-4251	NZ_CP027388
Escherichia Coli Strain 2013c-3264	NZ_CP027544
Escherichia Coli Strain 2013c-3277	NZ_CP027331
Escherichia Coli Strain 2013c-3342	NZ_CP027766
Escherichia Coli Strain 2013c-3513	NZ_CP027555
Escherichia Coli Strain 2013c-4187	NZ_CP027546
Escherichia Coli Strain 2014c-3051	NZ_CP027338
Escherichia Coli Strain 2014c-3655	NZ_CP027351
Escherichia Coli Strain 2015c-3163	NZ_CP027219
Escherichia Coli Strain 20r2r	CP062160

Escherichia Coli Strain 266917 2	NZ_CP026723
Escherichia Coli Strain 272	NZ_CP018239
Escherichia Coli Strain 28rc1	NZ_CP015020
Escherichia Coli Strain 319	NZ_CP018241
Escherichia Coli Strain 472	NZ_CP018245
Escherichia Coli Strain 9000	NZ_CP018252
Escherichia Coli Strain 95nr1	NZ_CP021339
Escherichia Coli Strain Atcc 43889	NZ_CP015853
Escherichia Coli Strain Ausmdu00002545	NZ_CP045975
Escherichia Coli Strain Ausmdu00014361	NZ_CP045827
Escherichia Coli Strain Cau16175	NZ_CP047378
Escherichia Coli Strain Cfsan027346	NZ_CP037945
Escherichia Coli Strain Cfsan027350	NZ_CP037941
Escherichia Coli Strain Cv261	NZ_CP040316
Escherichia Coli Strain Dsm 103246	NZ_CP019944
Escherichia Coli Strain E2855	NZ_AP018796
Escherichia Coli Strain Erl04-3476	NZ_CP032808
Escherichia Coli Strain Erl05-1306	NZ_CP032803
Escherichia Coli Strain Erl06-2442	NZ_CP032801
Escherichia Coli Strain Erl06-2497	NZ_CP032797
Escherichia Coli Strain Erl06-2503	NZ_CP032795
Escherichia Coli Strain F1 E4	NZ_CP040307
Escherichia Coli Strain Forc 044	NZ_CP016755
Escherichia Coli Strain Fsis11705876	NZ_CP035545
Escherichia Coli Strain Hb6	NZ_CP040305
Escherichia Coli Strain Ma11	NZ_CP040314

Escherichia Coli Strain Nadc 5570-86-24-6564 Isolate Wild Type	NZ_CP017251
Escherichia Coli Strain Nadc 5570-86-24-6565 Isolate Mutant	NZ_CP017249
Escherichia Coli Strain Nctc9112	NZ_LR134079
Escherichia Coli Strain Nzrm4169	NZ_CP032789
Escherichia Coli Strain Pa20	NZ_CP017669
Escherichia Coli Strain Pk8241	NZ_CP080139
Escherichia Coli Strain Rm13322	NZ_CP050498
Escherichia Coli Strain Rm19259	NZ_CP046527
Escherichia Coli Strain Scaid Wnd1-2021 (1-128)	NZ_CP082831
Escherichia Coli Strain Sj7	NZ_CP044315
Escherichia Coli Strain Srcc 1675	NZ_CP015023
Escherichia Coli Strain Stec2018-553	NZ_CP075665
Klebsiella Aerogenes 035	CP050067
Klebsiella Aerogenes 18-2341	CP049600
Klebsiella Aerogenes Ar 0007	CP024883
Klebsiella Aerogenes Ar 0018	CP024880
Klebsiella Aerogenes Ar 0062	CP026756
Klebsiella Aerogenes Ar0009	CP024885
Klebsiella Aerogenes Auh-Kam-9	CP048598
Klebsiella Aerogenes C9	CP042530
Klebsiella Aerogenes Cav1320	CP011574
Klebsiella Aerogenes Ea46506	CP070520
Klebsiella Aerogenes Fdaargos 1442	CP077293
Klebsiella Aerogenes Fdaargos 327	CP031756
Klebsiella Aerogenes Fdaargos 363	CP023963
Klebsiella Aerogenes G7	CP011539

Klebsiella Aerogenes Hnhf1	CP047669
Klebsiella Aerogenes Ka P10 L5 03.19	CP044214
Klebsiella Aerogenes Ka37751	CP041925
Klebsiella Aerogenes Kae3sp	CP082898
Klebsiella Aerogenes Lu2	CP035466
Klebsiella Aerogenes Ntt31xs Chromosome Ntt31xs-1	CP077429
Klebsiella Aerogenes Rhbstw-00898	CP056260
Klebsiella Aerogenes Rhbstw-00938	CP055904
Klebsiella Aerogenes Y1	CP045870
Klebsiella Aerogenes Y3	CP045869
Klebsiella Aerogenes Y6	CP045868
Klebsiella Africana 200023	CP084874
Klebsiella Africana Ff1003	CP059391
Klebsiella Grimontii 2481359	CP067380
Klebsiella Grimontii 5512.56	CP067391
Klebsiella Grimontii Kox 60	CP067433
Klebsiella Grimontii M5a1	CP063301
Klebsiella Grimontii Ny3	CP079754
Klebsiella Grimontii Rhbstw-00039	CP058180
Klebsiella Grimontii Rhbstw-00165	CP055364
Klebsiella Grimontii Rhbstw-00853	CP056150
Klebsiella Grimontii Rhbstw-00866	CP055309
Klebsiella Grimontii Ss141	CP044527
Klebsiella Huaxiensis Wchki090001	CP036175
Klebsiella Michiganensis 12084	CP072119
Klebsiella Michiganensis 53828cz	CP085764
Klebsiella Michiganensis 7525	CP065474

Klebsiella Michiganensis 8-1	CP089448
Klebsiella Michiganensis Akkl-001	CP060111
Klebsiella Michiganensis Ar375	CP029141
Klebsiella Michiganensis B106	CP067093
Klebsiella Michiganensis Bd-50-Km	CP061930
Klebsiella Michiganensis Bd177	CP048108
Klebsiella Michiganensis C52	CP042545
Klebsiella Michiganensis Ccri-24235	CP081351
Klebsiella Michiganensis Cz598	CP073305
Klebsiella Michiganensis Fdaargos 647	CP044109
Klebsiella Michiganensis Jnqh491	CP075881
Klebsiella Michiganensis K516	CP022348
Klebsiella Michiganensis K518	CP023185
Klebsiella Michiganensis K92	CP089315
Klebsiella Michiganensis Km41	CP090078
Klebsiella Michiganensis Kmfe267	CP071393
Klebsiella Michiganensis Knu07	CP041515
Klebsiella Michiganensis Ko 408	CP091470
Klebsiella Michiganensis Kox101	CP089407
Klebsiella Michiganensis Kox58	CP089395
Klebsiella Michiganensis Lds17	CP065338
Klebsiella Michiganensis M1	CP008841
Klebsiella Michiganensis M82255	CP035214
Klebsiella Michiganensis Rhb20-C02	CP058212
Klebsiella Michiganensis Rhbstw-00409	CP055325
Klebsiella Michiganensis Sb-24	CP073236
Klebsiella Michiganensis X2-1	CP051427

Klebsiella Oxytoca Ar 0147	CP020358
Klebsiella Oxytoca Cav1335	CP011618
Klebsiella Oxytoca Fdaargos 335	CP027426
Klebsiella Oxytoca Konih4	CP026269
Klebsiella Oxytoca Rhbstw-00493	CP056453
Klebsiella Pneumoniae Ba2275	CP053364
Klebsiella Pneumoniae F1	CP026130
Klebsiella Pneumoniae F138	CP026149
Klebsiella Pneumoniae F5	CP026132
Klebsiella Pneumoniae F77	CP026136
Klebsiella Pneumoniae Fdaargos 775	CP040993
Klebsiella Pneumoniae Frpdr	CP063759
Klebsiella Pneumoniae Kp1517	CP072463
Klebsiella Pneumoniae Kp18-1	CP082001
Klebsiella Pneumoniae Kp18-2079	CP048933
Klebsiella Pneumoniae Kp19-2029	CP047160
Klebsiella Pneumoniae Kp36	CP047192
Klebsiella Pneumoniae Kpc-2	CP078122
Klebsiella Pneumoniae P1428	CP017994
Klebsiella Pneumoniae Qd23	CP042858
Klebsiella Quasipneumoniae A708	CP026368
Klebsiella Quasipneumoniae G4584	CP034129
Klebsiella Quasipneumoniae Kqpf26	CP065838
Klebsiella Quasipneumoniae Ww-14a	CP080099
Klebsiella Quasipneumoniae Ydkl-002	CP068237
Klebsiella Quasivariicola 08a119	CP084768
Klebsiella Quasivariicola Kpn1705	CP022823

Klebsiella Variicola 342	CP000964
Klebsiella Variicola Dsm 15968	CP010523
Klebsiella Variicola F2r9	CP072130
Klebsiella Variicola Fh-1	CP054254
Klebsiella Variicola X39	CP018307
Mycobacterium Avium 104	CP000479
Mycobacterium Avium Dsm44156	CP046507
Mycobacterium Avium Fdaargos 1606	CP085978
Mycobacterium Avium Fdaargos 1607	CP085977
Mycobacterium Avium Hjw	CP028731
Mycobacterium Avium Rcad0278	CP016396
Mycobacterium Avium Subsp Hominissuis 101034	CP040247
Mycobacterium Avium Subsp Hominissuis 101115	CP040255
Mycobacterium Avium Subsp Hominissuis 101174	CP040250
Mycobacterium Avium Subsp Hominissuis Cam177	CP076851
Mycobacterium Avium Subsp Hominissuis Hp17	CP016818
Mycobacterium Avium Subsp Hominissuis Mac109	CP029332
Mycobacterium Avium Subsp Hominissuis Mah11	CP035744
Mycobacterium Avium Subsp Hominissuis Mc2 2500	CP036220
Mycobacterium Avium Subsp Hominissuis W14	CP060407
Mycobacterium Avium Subsp Hominissuis W9	CP060405
Mycobacterium Avium Subsp Homonissuis H87	CP018363
Mycobacterium Avium Subsp Paratuberculosis Dsm 44135	CP053068
Mycobacterium Avium Subsp Paratuberculosis Telford	CP033688
Mycobacterium Colombiense Cect 3035	CP020821
Mycobacterium Diernhoferi Atcc 19340	CP080332
Mycobacterium Frederiksbergense Lb 501t	CP038799

Mycobacterium Goodii X7b	CP012150
Mycobacterium Gordonae X7091	CP070973
Mycobacterium Grossiae Dsm 104744	CP043474
Mycobacterium Haemophilum Dsm 44634 Atcc 29548	CP011883
Mycobacterium Heraklionense Jcm 30995	CP080997
Mycobacterium Holsaticum Jcm 12374	CP080998
Mycobacterium Intracellulare Atcc 13950	CP003322
Mycobacterium Intracellulare Fdaargos 1564	CP085945
Mycobacterium Intracellulare Fdaargos 1611	CP089221
Mycobacterium Intracellulare Fdaargos 1612	CP089220
Mycobacterium Intracellulare Fdaargos 1616	CP089231
Mycobacterium Intracellulare Flac0133	CP023146
Mycobacterium Intracellulare Flac0181	CP023149
Mycobacterium Intracellulare Mott-02	CP003323
Mycobacterium Kansasii Fdaargos 1614	CP089218
Mycobacterium Kansasii Fdaargos 1615	CP089216
Mycobacterium Kansasii 10mk	CP019886
Mycobacterium Kansasii 11mk	CP019887
Mycobacterium Kansasii 4mk	CP019884
Mycobacterium Kansasii Atcc 12478	CP006835
Mycobacterium Kansasii Fdaargos 1534	CP084364
Mycobacterium Kubicae Jcm 13573	CP065047
Mycobacterium Kubicae Njh Mkub1	CP045081
Mycobacterium Kubicae Njh Mkub2	CP045075
Mycobacterium Leprae Mrhru-235-G	CP029543
Mycobacterium Malmoense Atcc 29571	CP080999
Mycobacterium Marinum 1218r	CP025779

Mycobacterium Marinum Ccug20998	CP024190
Mycobacterium Marinum Mma1	CP058277
Mycobacterium Marseillense Flac0026	CP023147
Mycobacterium Pallens Jcm 16370	CP080333
Mycobacterium Paragordonae 49061	CP025546
Mycobacterium Riyadhense Ntm	CP045092
Mycobacterium Senegalense Atcc 35796	CP081000
Mycobacterium Shigaense Scy	CP022927
Mycobacterium Shottsii M175	CP014860
Mycobacterium Tuberculosis 2-0029p6c4	CP041837
Mycobacterium Tuberculosis 2-0046p6c4	CP041833
Mycobacterium Tuberculosis 060827	CP058236
Mycobacterium Tuberculosis 1-0107p6c4	CP041853
Mycobacterium Tuberculosis 11502	CP070338
Mycobacterium Tuberculosis 120-26cao	CP071127
Mycobacterium Tuberculosis 2-0022p6c4	CP041840
Mycobacterium Tuberculosis 2-0023p6c4	CP041839
Mycobacterium Tuberculosis 267-47w148	CP071128
Mycobacterium Tuberculosis 3-0096p6c4	CP041827
Mycobacterium Tuberculosis 3-0124p6c4	CP041826
Mycobacterium Tuberculosis Beijing-Like-36918	CP017594
Mycobacterium Tuberculosis Beijing-Like-50148	CP017597
Mycobacterium Tuberculosis Cg20	CP072764
Mycobacterium Tuberculosis Cg21	CP072763
Mycobacterium Tuberculosis Cg23	CP072762
Mycobacterium Tuberculosis Cg24	CP072761
Mycobacterium Tuberculosis Dk9897	CP018778

Mycobacterium Tuberculosis F28	CP010330
Mycobacterium Tuberculosis Fdaargos 751	CP046308
Mycobacterium Tuberculosis Gg-111-10	CP025593
Mycobacterium Tuberculosis Gg-37-11	CP025598
Mycobacterium Tuberculosis Gg-90-10	CP025601
Mycobacterium Tuberculosis H107	CP019612
Mycobacterium Tuberculosis H37rv Cg	CP072765
Mycobacterium Tuberculosis H54	CP019610
Mycobacterium Tuberculosis H83	CP019611
Mycobacterium Tuberculosis Mt-0080	CP041207
Mycobacterium Tuberculosis Mtb1	CP020381
Mycobacterium Tuberculosis Mtb2	CP022014
Mycobacterium Tuberculosis Ocu901s S2 2s	CP018014
Mycobacterium Tuberculosis Rus B0	CP030093
Mycobacterium Tuberculosis Scaid 187.0	CP012506
Mycobacterium Tuberculosis Scaid 252-0	CP016888
Mycobacterium Tuberculosis Sea07010354p6c4	CP041813
Mycobacterium Tuberculosis Sit745-Eai1-Mys	CP046529
Mycobacterium Tuberculosis Tb282	CP017920
Mycobacterium Tuberculosis Tcdc11	CP046728
Mycobacterium Tuberculosis Wc059	CP022578
Mycobacterium Tuberculosis Wc078	CP022577
Mycobacterium Tuberculosis Zmc13-88	CP009101
Rhodococcus Aetherivorans Isolate Psbb011	NZ_CP069306
Rhodococcus Aetherivorans Strain Cbo21-1	NZ_CP088969
Rhodococcus Erythropolis Ccm2595	CP003761
Rhodococcus Erythropolis Pr4	AP008957

Rhodococcus Erythropolis R138	CP007255
Rhodococcus Erythropolis Strain Bg43	CP011295
Rhodococcus Erythropolis Strain Kb1	NZ_CP050124
Rhodococcus Erythropolis Strain R85	NZ_CP070870
Rhodococcus Erythropolis Strain X5	NZ_CP044284
Rhodococcus Fascians A21d2	NZ_CP049748
Rhodococcus Fascians A25f	NZ_CP049744
Rhodococcus Fascians D188	CP015235
Rhodococcus Fascians Strain Pbts 2	NZ_CP015220
Rhodococcus Globerulus Strain D757	CP079698
Rhodococcus Hoagii Jcm94-14	AP024181
Rhodococcus Hoagii Jcm94-16	AP024183
Rhodococcus Hoagii Jcm94-25	AP024185
Rhodococcus Hoagii Jcm94-27	AP024187
Rhodococcus Hoagii Jid03-27	AP024194
Rhodococcus Hoagii Jid03-46	AP024196
Rhodococcus Hoagii Jid03-56	AP024198
Rhodococcus Hoagii Strain Dsskp-R-001	CP027793
Rhodococcus Hoagii Strain Fdaargos 952	CP065594
Rhodococcus Hoagii Strain Jcm94-3	NZ_AP024192
Rhodococcus Hoagii Strain Jcm94-31	NZ_AP024189
Rhodococcus Hoagii Strain Wy	NZ_CP041647
Rhodococcus Koreensis Strain R85	NZ_CP070609
Rhodococcus Opacus Pd630	NZ_CP080954
Rhodococcus Opacus Strain 1cp	CP009111
Rhodococcus Opacus Strain Dsm 44186	CP082160
Rhodococcus Pyridinivorans Sb3094	CP006996

Rhodococcus Pyridinivorans Strain 5ap	NZ_CP063450
Rhodococcus Pyridinivorans Strain B403	NZ_CP066853
Rhodococcus Pyridinivorans Strain Gf3	NZ_CP022915
Rhodococcus Pyridinivorans Strain Tg9	CP022208
Rhodococcus Pyridinivorans Strain Yc-Jh2	NZ_CP050178
Rhodococcus Pyridinivorans Strain Yf3	NZ_CP040719
Rhodococcus Qingshengii Cs98	AP023172
Rhodococcus Qingshengii Strain 7b	CP063234
Rhodococcus Qingshengii Strain Cl-05	CP072108
Rhodococcus Qingshengii Strain Cx-1	CP054207
Rhodococcus Qingshengii Strain Rl1	CP042917
Rhodococcus Qingshengii Strain Tg-1	CP077417
Rhodococcus Rhodochrous Strain Atcc Baa870	NZ_CP032675
Rhodococcus Rhodochrous Strain Bx2	NZ_CP027557
Rhodococcus Ruber Strain C1	NZ_CP044211
Rhodococcus Ruber Strain P14	NZ_CP024315
Rhodococcus Ruber Strain R1	NZ_CP038030
Rhodococcus Ruber Strain Sd3	NZ_CP029146
Rhodococcus Ruber Strain Yc-Yt1	NZ_CP023714
Rhodococcus Ruber Strain Yyl	NZ_CP024890
Rhodococcus Sp. Djl-6-2	CP025959
Rhodococcus Triatomae Strain Dsm 44892	NZ_CP048814
Rhodococcus Triatomae Strain Dsm 44893	NZ_CP048813
Salmonella Bongori Cfsan000510	CP074233
Salmonella Bongori N268-08	CP006608
Salmonella Bongori Nctc 12419	FR877557
Salmonella Bongori Serovar 40 Z35 - Strain Cfsan001045	NZ_CP074592

Salmonella Bongori Serovar 48 Z41 -- Str. Rks3044	CP006692
Salmonella Bongori Serovar 48 Z81 - Strain 08-0158	CP053336
Salmonella Bongori Serovar 66 Z41 - Str. Sa19983605	CP022120
Salmonella Bongori Strain 04-0440	CP035676
Salmonella Bongori Strain 85-0051	CP053416
Salmonella Bongori Strain 92-0238	CP053417
Salmonella Bongori Strain Se40	CP067369
Salmonella Enterica Sehaa3795	AP020330
Salmonella Enterica Sesen3709	AP020332
Salmonella Enterica Strain 2011k-1440	NZ_CP053585
Salmonella Enterica Strain 2012k-0845	NZ_CP053579
Salmonella Enterica Strain 2014k-1020	NZ_CP053584
Salmonella Enterica Strain 85-0120	CP054715
Salmonella Enterica Strain 94-0093	NZ_CP053581
Salmonella Enterica Strain C629	NZ_CP015724
Salmonella Enterica Strain Cfsan033950	NZ_CP075141
Salmonella Enterica Strain Cfsan044865	NZ_CP075140
Salmonella Enterica Strain Cfsan044875	NZ_CP075138
Salmonella Enterica Strain Cfsan044885	NZ_CP075137
Salmonella Enterica Strain Cfsan044888	NZ_CP075135
Salmonella Enterica Strain Cfsan044925	NZ_CP075127
Salmonella Enterica Strain Cfsan044945	NZ_CP075125
Salmonella Enterica Strain Cfsan060804	CP075110
Salmonella Enterica Strain Cfsan060807	NZ_CP075109
Salmonella Enterica Strain Cfsan060808	NZ_CP075108
Salmonella Enterica Strain Cfsan060809	NZ_CP075107
Salmonella Enterica Strain Cfsan064033	NZ_CP028172

Salmonella Enterica Strain Cfsan064034	NZ_CP028169
Salmonella Enterica Strain Cfsan064276	CP075106
Salmonella Enterica Strain Cfsan096147	NZ_CP044257
Salmonella Enterica Strain Da34821	NZ_CP029567
Salmonella Enterica Strain Fdaargos 1066	NZ_CP066047
Salmonella Enterica Strain Fdaargos 1067	NZ_CP066009
Salmonella Enterica Strain Fdaargos 1271	NZ_CP069518
Salmonella Enterica Strain Fdaargos 687	NZ_CP046283
Salmonella Enterica Strain Fdaargos 688	NZ_CP046280
Salmonella Enterica Strain Fdaargos 70	NZ_CP026052
Salmonella Enterica Strain Fdaargos 707	NZ_CP046279
Salmonella Enterica Strain Fdaargos 708	NZ_CP046278
Salmonella Enterica Strain Fdaargos 710	NZ_CP046277
Salmonella Enterica Strain Fdaargos 712	NZ_CP046291
Salmonella Enterica Strain Fdaargos 765	NZ_CP041011
Salmonella Enterica Strain Fdaargos 768	NZ_CP041005
Salmonella Enterica Strain Fdaargos 878	CP065718
Salmonella Enterica Strain Fdaargos 928	NZ_CP065639
Salmonella Enterica Strain Forc 019	NZ_CP012396
Salmonella Enterica Strain Forc 038	NZ_CP015574
Salmonella Enterica Strain Forc 051	NZ_CP017232
Salmonella Enterica Strain Forc 056	NZ_CP017177
Salmonella Enterica Strain Forc 074	NZ_CP023436
Salmonella Enterica Strain Forc 078	NZ_CP026713
Salmonella Enterica Strain Fsw0104	NZ_CP037894
Salmonella Enterica Strain Gsj 2017-Sal.-009	NZ_CP050833
Salmonella Enterica Strain Gsj2016-Sal.-018	NZ_CP069166

Salmonella Enterica Strain Gx1006	NZ_CP060585
Salmonella Enterica Strain K Sa184	NZ_CP061159
Salmonella Enterica Strain Lhica E3	NZ_CP079839
Salmonella Enterica Strain Lt2	CP014051
Salmonella Enterica Strain Mac15	NZ_CP030749
Salmonella Enterica Strain Mfds1004024	NZ_CP025745
Salmonella Enterica Strain Mfds1004839	NZ_CP026569
Salmonella Enterica Strain No75	NZ_CP075372
Salmonella Enterica Strain Osf005645	NZ_CP040380
Salmonella Enterica Strain Qh	NZ_CP043773
Salmonella Enterica Strain S146	NZ_CP077662
Salmonella Enterica Strain S44712	NZ_CP035917
Salmonella Enterica Strain S61394	NZ_CP035915
Salmonella Enterica Strain S639	NZ_CP089207
Salmonella Enterica Strain S90	NZ_CP077670
Salmonella Enterica Strain Sa19992307	NZ_CP030207
Salmonella Enterica Strain Sa20021456	NZ_CP030219
Salmonella Enterica Strain Sa20025921	NZ_CP030214
Salmonella Enterica Strain Sa20030575	NZ_CP030181
Salmonella Enterica Strain Sa20031245	NZ_CP030235
Salmonella Enterica Strain Sa20041605	NZ_CP030225
Salmonella Enterica Strain Sa20043041	NZ_CP030231
Salmonella Enterica Strain Sa20044414	CP030209
Salmonella Enterica Strain Sa20051401	NZ_CP030196
Salmonella Enterica Strain Sa20051528	NZ_CP030211
Salmonella Enterica Strain Sa20052327	NZ_CP030202
Salmonella Enterica Strain Sa20055162	NZ_CP030238

Salmonella Enterica Strain Sa20075157	NZ_CP030217
Salmonella Enterica Strain Sa20080453	CP030194
Salmonella Enterica Strain Sa20083039	NZ_CP030223
Salmonella Enterica Strain Sa20083530	NZ_CP030203
Salmonella Enterica Strain Sa20094620	NZ_CP030185
Salmonella Enterica Strain Sa20100201	NZ_CP030180
Salmonella Enterica Strain Sa20101045	NZ_CP030233
Salmonella Enterica Strain Sa20104250	NZ_CP030190
Salmonella Enterica Strain Scsw714	NZ_CP051213
Salmonella Enterica Strain Sg1722-1	NZ_CP081187
Salmonella Enterica Strain Slr1 7627	NZ_CP060517
Salmonella Enterica Strain Slr1 7697	NZ_CP060508
Salmonella Enterica Strain Slr1 7966	NZ_CP060512
Salmonella Enterica Strain Slr1 8094	NZ_CP060515
Salmonella Enterica Strain Slr1 8245	NZ_CP080091
Salmonella Enterica Strain Slr1 8250	NZ_CP060522
Salmonella Enterica Strain Sp	NZ_CP077668
Salmonella Enterica Strain Src27	NZ_CP058807
Salmonella Enterica Strain Szl 30	NZ_CP085981
Salmonella Enterica Strain Szl 31	NZ_CP085983
Salmonella Enterica Strain Szl 38	NZ_CP085987
Yersinia Pestis 1412	CP006783
Yersinia Pestis 1413	CP006762
Yersinia Pestis 1522	CP006758
Yersinia Pestis 2944	CP006792
Yersinia Pestis 3067	CP006754
Yersinia Pestis 3770	NZ_CP006751

Yersinia Pestis 8787	NZ_CP006748
Yersinia Pestis A1122	CP002956
Yersinia Pestis Angola	CP000901
Yersinia Pestis Antiqua	CP000308
Yersinia Pestis Co92	AL590842
Yersinia Pestis D182038	CP001589
Yersinia Pestis Ev Niieg	CP065918
Yersinia Pestis Kim10+	AE009952
Yersinia Pestis Pestoides F	NZ_CP009715
Yersinia Pestis Str. Pestoides B	NZ_CP010023
Yersinia Pestis Strain 14d	NZ_CP063303
Yersinia Pestis Strain 195-P	CP019708
Yersinia Pestis Strain C-781	NZ_CP064117
Yersinia Pestis Strain C-783	NZ_CP064118
Yersinia Pestis Strain C-792	NZ_CP064119
Yersinia Pestis Strain C-830	CP064127
Yersinia Pestis Strain Cadman	NZ_CP016273
Yersinia Pestis Strain El Dorado	NZ_CP009785
Yersinia Pestis Strain Fdaargos 601	NZ_CP033699
Yersinia Pestis Strain Fdaargos 602	NZ_CP033696
Yersinia Pestis Strain Fdaargos 603	NZ_CP033690
Yersinia Pestis Strain Harbin35	NZ_CP009704
Yersinia Pestis Strain I-1252	NZ_CP064126
Yersinia Pestis Strain Java9	NZ_CP009996
Yersinia Pestis Strain Km 567	CP064120
Yersinia Pestis Strain M-1482	NZ_CP064121
Yersinia Pestis Strain M-1974	NZ_CP064124

Yersinia Pestis Strain M2029	NZ_CP064123
Yersinia Pestis Strain M2085	NZ_CP064125
Yersinia Pestis Strain M2086	NZ_CP064128
Yersinia Pestis Strain Nicholisk 41	NZ_CP009991
Yersinia Pestis Strain Pbm19	NZ_CP009492
Yersinia Pestis Strain S19960127	NZ_CP045640
Yersinia Pestis Strain Scpm-O-B-5935 (I-1996)	NZ_CP045154
Yersinia Pestis Strain Scpm-O-B-5942 (I-2638)	NZ_CP045258
Yersinia Pestis Strain Scpm-O-B-6291 (C-25)	NZ_CP045163
Yersinia Pestis Strain Scpm-O-B-6899 (231)	NZ_CP045145
Yersinia Pestis Strain Scpm-O-Dna-18 (I-3113)	NZ_CP045149
Yersinia Pestis Strain Shasta	NZ_CP009723
Yersinia Pestis Subsp. Pestis Bv. Medievalis	CP045158
Yersinia Pestis Z176003	CP001593