

**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS**  
ESCOLA POLITÉCNICA E DE ARTES / ENGENHARIA DA COMPUTAÇÃO  
Trabalho Conclusão de Curso II

**Gabriel Henrique Lima Picouto**

**BIOINFORMATICA E A UTILIZAÇÃO DA GENÉTICA NA AGRICULTURA**

Trabalho de Conclusão de Curso II como parte dos requisitos para obtenção do título de bacharel em Engenharia da Computação apresentado à Pontifícia Universidade Católica de Goiás.

**BANCA EXAMINADORA:**

Prof. Dr. Antônio Marcos de Melo Medeiros – Orientador. ENG-PUC Goiás.  
Prof. Dr. Bruno Quirino de Oliveira.  
Prof. Dr. Marcos Antônio de Sousa

Goiânia, 15 de dezembro de 2023.

# BIOINFORMATICA E A UTILIZAÇÃO DA GENÉTICA NA AGRICULTURA

Picouto. G. H. L, Medeiros A. M. M, Oliveira B. Q, Souza A. S

Escola politécnica e de artes, PUC-GOÍÁS.

**Resumo** –A bioinformática é uma área que combina conhecimentos de computação e ciências biológicas para analisar sequências de DNA, proteínas e RNA. Sua principal tarefa é lidar com grande volume de dados gerados pela biologia molecular e genômica, especialmente na era digital, onde o sequenciamento de biomoléculas se tornou mais rápido, acessível e gera grandes quantidades de dados. A bioinformática utiliza métodos computacionais de alto desempenho para analisar esses dados e obter respostas e soluções como o desenvolvimento de um Sistema de identificação de marcadores moleculares, com o objetivo de auxiliar a seleção de plantas com certas características desejáveis, como por exemplo, maior resistência a seca ou a doenças, maior produtividade

**Palavras.** bioinformática, agricultura, genoma

## BIOINFORMATICS AND THE USE OF GENETICS IN AGRICULTURE

**Abstract** - Bioinformatics is an area that combines knowledge of computation and biological sciences to analyze sequences of DNA, proteins and RNA. Its main task is to deal with the large volume of data generated by molecular biology and genomics, especially in the digital age, where the sequencing of biomolecules has become faster, more accessible and generates large amounts of data. Bioinformatics uses high-performance computational methods to analyze these data and obtain answers and solutions, such as the development of a molecular marker identification system, with the aim of helping to select plants with certain desirable characteristics, such as greater resistance to drought or disease, higher productivity.

**Keyword :** bioinformatics, agriculture, genome

### I. INTRODUÇÃO

A Bioinformática atua fazendo a união de dois campos de conhecimento meio distantes entre si, sendo eles, a Computação e as Ciências Biológicas. Aplicando técnicas e conhecimentos da computação em áreas específicas das Ciências Biológicas, como Genética, Ciências Molecular, Genômica. Assim a bioinformática não é considerada como uma nova área da ciência propriamente dita, mas sim, como um programa metodológico de pesquisas que aplica técnicas e conhecimentos estudados em uma área em uma outra para alcançar novos resultados [1].

O foco principal da Bioinformática é analisar através de métodos computacionais sequências de DNA, proteínas e RNA [2]. Com o aumento no volume de dados para serem estudados e analisados da biologia molecular e genômica, principalmente agora na era digital onde são gerados mais rápidos e em maior quantidade e mais acessíveis (Big-Data), que são gerados com as tecnologias de sequenciamento. A bioinformática se tornou a

principal responsável pela análise deles utilizando métodos computacionais de alto desempenho para obter respostas e soluções para questionamentos complexos de estudos que envolvam sequencias de biomoléculas [3].

Para poder armazenar todos esses dados que são gerados foram criados os Bancos de Dados Biológicos, podendo ser de acesso público ou privado, onde podem ser encontrados informações de várias áreas de pesquisa relacionadas a Genômica, Molecular, Proteica entre outras. Um dos principais é o GenBank, sendo um banco de dados de sequenciamento de DNA. Existem também o EMBL (European Molecular Biology Laboratory), o NCBI (National Center for Biotechnology Information) [1]. No Brasil existem laboratórios que são utilizados em pesquisas voltadas para a Bioinformática, sendo alguns deles: Projeto Genoma Brasileiro [1] e o Laboratório Multiusuário de Bioinformática (LMB) da Embrapa [3].

Existiam duas formas de fazer experimentos biológicos antes da bioinformática: o *in vivo*, que utiliza seres vivos, e o *in vitro*, utiliza um sistema artificial [5]. Podemos dividir os estudos e pesquisas voltadas para a bioinformática, que é *in silico*, em três tipos de princípios: ‘Tijolo’ onde o foco é produzir formas e ferramentas de analisar e entender as sequencias de genoma. ‘Peneira’ que é usada para minerar os dados de genomas, principalmente quando se tem fragmentos de genomas diferentes em outras pesquisas relacionadas. ‘Lupa’ é onde todos os dados obtidos são analisados, é feito experimentos, criação de hipóteses utilizando a computação com isso e possível confirmá-la ou negá-la [2].

De fato a Bioinformática ajudou muito a evoluir questões de pesquisas desde as básicas a aplicada nas últimas décadas. Desde o setor farmacêutico com a manipulação de moléculas para a fabricação de remédio, na medicina onde podem ser desenvolvidas tratamentos terapêuticos e preventivos contra doenças específicas, na agricultura com modificação de marcadores genéticos para criar plantas resistentes a certas condições climáticas. Isso mostra o impacto e alcance da bioinformática, quando junta ferramentas computacionais nas pesquisas biológicas [4].

O objetivo desse trabalho é mostrar a atuação e melhora que o uso da Bioinformática causou em duas áreas muito importantes na sociedade: na saúde humana e na agricultura. Demonstrar de forma pratica a utilização dela na agricultura aplicando o Random Florest em um sistema de identificação de marcadores moleculares de plantas.

### II. FUNDAMENTAÇÃO TEÓRICA

Este campo interdisciplinar, que une biologia e ciência da computação, demonstrou seu impacto vital na genética, especialmente nas áreas de saúde e agricultura. Em saúde, a bioinformática impulsiona a medicina personalizada, permitindo diagnósticos mais precisos e tratamentos adaptados às nuances genômicas individuais. Na agricultura, essa disciplina desempenha um papel crucial na criação de cultivos mais robustos e adaptáveis, promovendo avanços na segurança alimentar.

#### A. Medicina, Genética e Saúde

Na interseção da medicina, genética e saúde, a bioinformática emerge como uma disciplina catalisadora, redefinindo os paradigmas de pesquisa e aplicação prática. Ao desbravar os vastos terrenos do código genético, a bioinformática torna-se a chave mestra que desbloqueia os segredos intrincados da genômica, proporcionando visões cruciais para avanços significativos. Desde a personalização de tratamentos médicos até a identificação de marcadores genéticos associados a doenças, sua influência abrangente transcende os limites tradicionais, prometendo moldar o futuro da medicina, genética e saúde.

a. Projeto Genoma Humano

Para tratar doenças que são originadas a partir de alterações genéticas, é necessário estudar e compreender o gene humano, que possui um genoma com cerca de 3 bilhões de pares de base, equivalente a 750 megabytes de memória de computador. Com a intenção de mapear e sequenciar o genoma, com a finalidade de identificar todas as suas estruturas e possíveis mutações, que surgiu em 1990, o mais importante projeto de bioinformática, o Projeto Genoma Humano (PGH), possuindo certas propostas para alcançar esse objetivo, mostradas no Quadro 1 [6].

Quadro 1: Proposta PGH, Projeto Genoma Humano.

Mapa genético	Marcadores a cada 2 – 5 cM
Mapa físico	Marcadores a cada 100 kb
Sequenciamento do genoma todo	Melhorar as técnicas de sequenciamento e desenvolver novas tecnologias, bem como reduzir os custos
Organismos modelos	Mapeamento e sequenciamento de alguns organismos: camundongo, Escherichia coli, levedura, Caenorhaditis elegans, Drosophila
Informática	Desenvolver programas e bancos de dados contendo os mapas genéticos e físicos do genoma humano e outros
Problemas éticos, legais e sociais	Desenvolver programas para estudar as implicações éticas (e outras) em decorrência do PGH
Transferência de tecnologia	Desenvolver tecnologia visando a automação do sequenciamento de DNA e mapeamento: estimular a transferência da tecnologia para a indústria e a comunidade médica

Com o avanço do PGH, houve grandes impactos nos métodos de identificação de genes que são a análise da proteína e o mapeamento dos genes em um cromossomo. Outro impacto é no diagnóstico de doenças genéticas, que tem como alvo os testes genéticos, que podem identificar pessoas que possuem genes com patógenos de doenças, com isso ter informações de doenças futuras não só do indivíduo, mas também de parentes, podem assim começar um tratamento preventivo [6].

Projeto Genoma Humano que iniciou o surgimento dos bancos de dados biológicos abertos para o compartilhamento das informações coletadas dos genes humanos, com isso mostrou o impacto do uso da informática no meio biológico [6].

o Digitalização do DNA e a NGS

Para que o DNA humano seja digitalizado é necessário que passa por vários processos que vão desde etapas bioquímicas a utilização de softwares para a transformação do DNA em dados eletrônicos formados por uma sequência de A, T, C, G. São utilizados algoritmos para sequenciar os fragmentos gerados na digitalização, o arquivo gerado pode ser em dois padrões universais [8]:

- FASTA: É composto por sequências de nucleótidos de DNA, cada sequência possui 2 linhas, a primeira é o identificador/cabeçalho e a segunda contém os dados/bases lidos do DNA, mostrado na Figura 1.
- FASTQ: Possui a mesma estrutura que o FASTA, com o acréscimo de uma linha contendo informações sobre a qualidade da sequência analisada, importantes para detecção de mutações, mostrado na Figura 3.

Figura 1 : FASTA [7]

```
>Taxon1
CCTGCGGAAGATCGGCACTAGAAATAGCCAGAACCCTTTCTCTGAGGCTTCCGGCTTCCC
>Taxon2
CCATCGGTAGCGCATCCTTAGTCCAATTAAGTCCCTATCCAGGCGCTCCGCCGAAGGTCT
>Taxon3
CCACCTCGTGGTATGGCTAGGCATTAGGAAACCGGAGAACGCTTCAGACCAGCCCGGAC
```

Figura 2: FASTQ [7]

```
@HWI-ST999:102:D1N6AACXX:1:1101:1235:1936 1:N:0: (IDENTIFICADOR)
ATGTCTCCTGGACCCCTCTGTGCCAAGCTCCTCATGCATCCTCTCAGCAACTGTCTGTAGCTGAGGCTCA
CTGACTACCAGCTGCAG (SEQUENCIA)
+
1:DAADDDF<B<\AGF=FGIEHCCD9DG=1E9?D>CF@HNG?B<GEBGHCG;;CDB8==C@@>GII@5?
A?@B>CEDCFCC;;?CCAC (QUALITY SCORE)
```

A Next-generation sequencing (NGS) vem utilizando uma nova tecnologia de sequenciadores que podem fazer sequenciamentos de genomas em grandes quantidades, com essa velocidade é esperado que as identificações de novos genes de doenças sejam descobertas muito mais rápidos. Com esse aumento de novos dados, novos procedimentos desenvolvidos pela bioinformática se tornam essências para facilitar a análise médica de dados genômicos que são encontrados em cada paciente. O próximo avanço é trazer essa tecnologia de sequenciamento para hospitais para serem usados em exames clínicos de rotina, principalmente voltados para doenças genéticas [8].

b. Medicina

A bioinformática vem ajudando muito a medicina em vários quesitos, desde a utilização de equipamentos e softwares especializados é possível identificar moléculas que podem servir como matéria-prima na produção de remédios, facilitar a detecção de novos agentes parasitas e com isso facilitar o diagnóstico e na epidemiologia de inúmeras doenças [4].

Com o tempo os dados genéticos de pacientes vão se tornar muito importantes na hora que o médico vai decidir tratamentos e soluções clínicas, surgindo assim a Medicina Personalizada, que com os dados genéticos do paciente pode usar melhores ferramentas e estratégias para decidir o melhor remédio, mais adequado, evitando efeitos de alergia, ter a noção exata da dose a ser utilizada a partir de biomarcadores farmacogenômicos dos pacientes [8].

Muito discutida pela sociedade científica, a Técnica de Terapia Genética CRISPR, que pode ser utilizada para a manipulação genética de forma barata e rápida, assim podendo evitar que possa por exemplo nasçam com o gene do HIV. Essa técnica se baseia no sistema de defesa de bactérias que pega um

pedaço do DNA do agente patogênico e deixa-o inativo dentro do seu próprio DNA de forma a criar uma resistência contra ele. Com isso, teste em laboratório mostraram que essa técnica pode ser utilizada para modificar a genética humana para evitar doenças [9].

### B. Agricultura

A bioinformática esta muito presente também na agricultura, com o uso de ferramentas computacionais para coletar dados e o sequenciamento de genomas de plantas tem como objetivo tornar as colheitas mais resistentes a condições climáticas extremas, doenças e aumentar a produção. Uma das maneiras estudadas para realizar essa melhora na produção agrícola é melhora nutricional do solo, estudando as comunidades microbianas presentes nele [10]. A NGS é utilizada para fazer o sequenciamento dessas comunidades microbianas ao fazer o sequenciamento do gene 16S RNAr de bactérias, podendo identificar genes que agem como inseticidas e resistentes a pragas [11].

Como as plantas utilizam de certas comunidades microbianas como fonte de defesa para agentes patógenos, como fertilizantes e controle de minerais, o estudo sobre o solo e a análise de micro-organismos que vivem neles é de grande importância para gerar resultados positivos na agricultura [11].

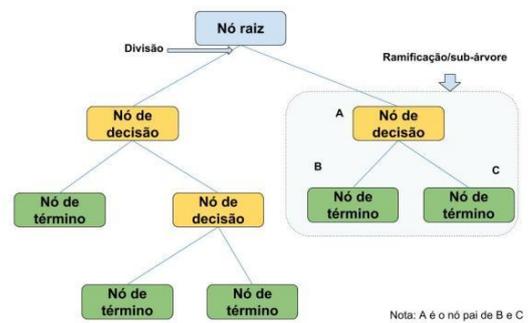
NGS utilizado na agricultura vem com o objetivo de identificar, manipular genes para condições fenotípicas específicas e reprodução por seleção genômica baseada em marcadores genéticos, são soluções que a pesquisa genômica encontra para agricultura a fim de desenvolver sementes transgênicas para uma produção sustentável, sem utilização de pesticidas e aumento de produtividade [12].

Devido a alta taxa de dados genômicos, o desafio no desenvolvimento de ferramentas computacionais aumenta também, a grandes bancos de dados biológicos espalhados pelo mundo, alguns especializados em genoma de plantas como Índice de Genes no Instituto de Pesquisa Genômica (TIGR), mas o mercado necessita de ferramentas também que ajudem a obter esses dados ou que utilizem eles em outras pesquisas como identificar simbiontes de plantas e organismos causadores de doenças como bactérias e fungos [10].

### C. Random Florest

O sistema utiliza o método de aprendizagem o Random Florest. Primeiramente, uma árvore de decisão é um modelo de aprendizado de máquina que representa decisões e suas possíveis consequências em forma de uma árvore. Cada nó interno da árvore representa uma decisão baseada em uma característica específica, e cada folha representa o resultado ou a classe prevista. Durante o treinamento, o algoritmo de árvore de decisão escolhe a melhor característica para dividir os dados com base em critérios como ganho de informação. Conforme é mostrado na figura 3.

Figura 3: Funcionamento de uma árvore de decisão



A Random Forest é um aprendizado em conjunto baseado em árvores de decisão. Ela constrói várias árvores de decisão durante o treinamento e combina suas previsões para obter um resultado mais robusto e geral. Cada árvore na floresta é treinada em uma subamostra aleatória dos dados de treinamento e, frequentemente, com um subconjunto aleatório das características. Conforme é mostrado na figura 4.

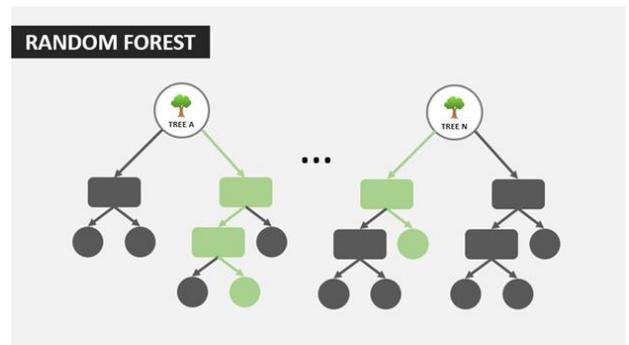


Figura 4 : Várias árvores de decisão vira um Random Forest

## III. Materiais e métodos

Para o desenvolvimento de um sistema de identificação de marcadores moleculares, para auxiliar na seleção de soja que possuem características que são desejáveis para as condições climáticas de uma certa região com objetivos de ter maior produtividade, resistência a doenças e pragas entre outras características.

### A. Banco de dados:

O objetivo inicial era utilizar dados genômicos de soja obtidos a partir de banco de dados biológicos públicos e de empresas que disponibilizam essas informações. Mas os dados disponíveis em sua grande maioria, são informações gerais da planta, não informando características específicas delas, como resistência maior a climas secos por exemplo.

Com isso foi realizado pesquisas das sojas utilizadas na região de Goiás e o clima da região, foi criado uma base de dados, para treinar a IA do sistema, utilizando 5 características genéticas sendo elas: resistência ao calor, tamanho, taxa de crescimento, adaptabilidade ao solo, produção de semente com os scores indo de 0 a 1. Também 2 características do clima sendo: radiação e umidade com scores de 1 a 100 e a informação se a planta sobreviveu ou não, como mostra a tabela 1 abaixo:

Tabela 1: Base de dados

genetica1	genetica2	genetica3	genetica4	genetica5	clima1	clima2	sobreviveu
0,4	0,2	0,9	0,3	0,7	34	5	1
0,6	0,8	0,2	0,8	0,1	47	6	0
0,3	0,7	0,5	0,4	0,5	52	4	1
0,5	0,1	0,6	0,7	0,2	39	3	1
0,8	0,9	0,1	0,2	0,8	63	7	0
0,4	0,4	0,3	0,5	0,9	21	4	1
0,7	0,5	0,8	0,1	0,4	49	2	1
0,2	0,3	0,4	0,9	0,6	57	1	0
0,9	0,6	0,7	0,6	0,3	29	5	1
0,1	0,8	0,2	0,5	0,5	65	3	0
0,5	0,3	0,9	0,7	0,8	24	2	1
0,3	0,9	0,6	0,8	0,2	46	4	0
0,4	0,7	0,1	0,2	0,7	32	6	1
0,8	0,5	0,4	0,6	0,1	43	8	0
0,6	0,2	0,3	0,4	0,9	38	3	1
0,7	0,1	0,5	0,1	0,3	41	7	0
0,2	0,6	0,8	0,3	0,4	27	5	1
0,9	0,4	0,7	0,5	0,6	35	2	1
0,1	0,5	0,2	0,9	0,8	37	6	0
0,3	0,8	0,4	0,7	0,2	51	4	1

Cada característica é representada por um Identificador numérico, que no código para ser referencia para uma característica a escolha do produtor, o mesmo se aplica as condições climáticas. A informação se a planta sobreviveu é representada pelo número 1 e não sobreviveu pelo número 0. Algumas características foram de sojas já utilizadas nas condições climáticas impostas, e outras foram obtidas a partir de condições simuladas.

#### B. Desenvolvimento do Sistema com IA:

A linguagem de programação utilizada para o desenvolvimento do código do sistema foi Python devido ao fato de já ser bastante utilizada na área de bioinformática pois possui uma flexibilidade e disponibilidade de bibliotecas especializadas, uma das principais bibliotecas utilizadas foi a Biopython.

Como a bioinformática trabalha com um grande volume de dados para serem analisados, o sistema desenvolvido utiliza uma IA, que é treinada com a base de dados fornecida. Quando um arquivo que contém uma nova sequência de características de uma soja e do clima, o sistema analisa para ver se ela vai sobreviver ou não, baseado nos dados em que ela foi treinada.

A Random Forest é então criada e treinada utilizando o RandomForestClassifier. Durante o treinamento, diversas árvores de decisão são construídas. Cada árvore é treinada em uma subamostra aleatória dos dados de treinamento e com uma seleção aleatória de características em cada divisão de nó. Esse processo visa reduzir o overfitting e aumentar a capacidade de generalização do modelo para novos dados.

A baixo segue um fluxograma de funcionamento do código

Início:

- Processo: Ler os dados de treinamento da base de dados
- Processo: Separar as características das plantas (X) da variável resposta (y)
- Processo: Separar os dados em conjuntos de treinamento e teste
- Processo: Criar um modelo de floresta aleatória
- Processo: Treinar o modelo no conjunto de treinamento
- Processo: Avaliar o desempenho do modelo no conjunto de teste

- Processo: Ler os dados da nova planta de um arquivo txt
- Processo: Fazer uma previsão de sobrevivência para a nova planta
- Decisão: Se a previsão for 1, ir para “Planta ideal”
- Decisão: Se a previsão for 0, ir para “Planta não ideal”
- Planta ideal
- Processo: Imprimir “Esta planta é ideal para as condições fornecidas e provavelmente sobreviverá.”
- Planta não ideal
- Processo: Imprimir “Esta planta não é ideal para as condições fornecidas. Procurando uma alternativa...”
- Processo: Encontrar plantas similares que sobreviveram
- Encontrando plantas similares que sobreviveram
- Processo: Encontrar plantas com as mesmas características genéticas da nova planta que sobreviveram
- Decisão: Se encontrar plantas similares que sobreviveram, ir para “Recomendar planta similar”
- Decisão: Se não encontrar plantas similares que sobreviveram, ir para “Não foi encontrada nenhuma planta similar adequada nas condições fornecidas”
- Recomendar planta similar
- Processo: Imprimir “Uma planta com maiores chances de sobrevivência tem as seguintes características genéticas:”
- Processo: Imprimir as características genéticas da primeira planta similar que sobreviveu
- Não foi encontrada nenhuma planta similar adequada nas condições fornecidas
- Processo: Imprimir “Não foi encontrada nenhuma planta similar adequada nas condições fornecidas.”

A complexidade computacional do treinamento de uma Random Forest depende principalmente do número de árvores na floresta (M), do número de amostras no conjunto de treinamento (N), e do número médio de features (K) em cada árvore. A complexidade geral de tempo para treinar uma Random Forest é frequentemente aproximada por, equação 1:

$$O(M * N * \log(K)) \quad (1)$$

Sendo:

M: número de árvores na floresta.

N: número de amostras no conjunto de treinamento.

K: número médio de features em cada árvore.

Essa complexidade reflete o custo de treinar cada árvore individualmente e a construção do conjunto. A complexidade é dominada pelo treinamento de cada árvore, que é  $O(N * \log(K))$ . O fator adicional de M reflete o número de árvores na floresta.

Vale ressaltar que, apesar da complexidade computacional relativamente alta, as Random Forests são geralmente eficientes na prática e podem lidar com conjuntos de dados relativamente grandes. A escolha do número de árvores e outros parâmetros pode influenciar o desempenho e a eficiência do treinamento.

Após passar por essa análise, o sistema dá uma saída,

conforme na figura 5 e figura 6, com as informações sobre a nova soja analisada. Contendo a probabilidade de sobrevivência, a probabilidade de não sobrevivência. Caso a soja não sobreviva, o sistema procura uma planta semelhante na base de dados, caso haja alguma ele indica a sequência de características dela, se não for encontrado, é informado que não foi encontrado nenhuma planta similar adequada as condições fornecidas. Uma soja é ideal para a sobrevivência nas condições fornecidas a taxa de sobrevivência deve ser maior que 50%.

Figura 5: Dados da nova soja analisada

```
Esta planta é ideal para as condições fornecidas e provavelmente sobreviverá.
Probabilidade de sobrevivência: 89,00%
Probabilidade de não sobrevivência: 11,00%
```

Figura 6 : Dados da nova soja analisada

```
Esta planta não é ideal para as condições fornecidas. Procurando uma alternativa...
Não foi encontrada nenhuma planta similar adequada nas condições fornecidas.
Probabilidade de sobrevivência: 50,00%
Probabilidade de não sobrevivência: 50,00%
```

Na Random Forest, a probabilidade de sobrevivência ou não sobrevivência é agregada a partir das previsões de todas as árvores na floresta. Cada árvore individual faz sua própria previsão com base nas características da planta, e a probabilidade final é determinada pela média (ou votação, dependendo do contexto) das probabilidades de todas as árvores.

Quando você chama o método `predict\_proba` no modelo de Random Forest, ele retorna uma matriz de probabilidades para cada classe. Essas probabilidades são agregadas a partir de todas as árvores na floresta.

Portanto, a probabilidade final é uma média ponderada das probabilidades estimadas por cada árvore. O número de árvores na floresta contribui para essa média ponderada. Se o modelo foi treinado com, por exemplo, 100 árvores, as previsões de cada árvore são combinadas para formar a previsão final, levando em consideração a contribuição de todas as árvores na floresta.

Essa abordagem de combinar as previsões de várias árvores é o que confere robustez e estabilidade ao modelo, reduzindo o risco de *overfitting* a padrões específicos nos dados de treinamento.

#### IV. Resultados

Após o treinamento, a precisão do modelo é avaliada usando dados de teste, medindo quão bem ele se ajusta a novos dados. Em seguida, o modelo é aplicado a dados de uma nova planta para prever se ela é adequada às condições fornecidas. O código apresenta uma mensagem indicando a adequação da planta e, caso não seja ideal, procura por plantas similares que sobreviveram, sugerindo alternativas.

Além disso, o código fornece uma análise de probabilidade de sobrevivência e não sobrevivência para a nova planta. Essa abordagem oferece não apenas previsões, mas também *insights* sobre a probabilidade associada a cada classe, permitindo uma avaliação mais detalhada das decisões do modelo.

O sistema usa o processo de treinamento e aplicação de um modelo de Random Forest para a classificação de plantas, destacando sua capacidade de oferecer previsões precisas e percepções sobre plantas similares que podem ser consideradas como alternativas em condições semelhantes.

Ao inserir novos dados de sojas novas para serem analisados conseguimos esses resultados, conforme é mostrado no quadro 2, quadro 3, quadro 4, quadro 5, quadro 6, quadro 7 e

quadro 8.

Quadro 2: Análise soja1

Soja 1	
Planta é ideal para as condições fornecidas?	Sim, a planta é ideal para as condições fornecidas
Taxa de sobrevivência	89,00%
Taxa de não sobrevivência	11,00%
Caso não sobreviva, foi encontrado soja similar na base de dados?	-----

Quadro 3: Análise soja2

Soja 2	
Planta é ideal para as condições fornecidas?	Sim, a planta é ideal para as condições fornecidas
Taxa de sobrevivência	95,00%
Taxa de não sobrevivência	5,00%
Caso não sobreviva, foi encontrado soja similar na base de dados?	-----

Quadro 4: Análise soja3

Soja 3	
Planta é ideal para as condições fornecidas?	Não, a planta não é ideal para as condições fornecidas
Taxa de sobrevivência	50,00%
Taxa de não sobrevivência	50,00%
Caso não sobreviva, foi encontrado soja similar na base de dados?	Não foi encontrado nenhuma planta similar adequada nas condições fornecidas

Quadro 5 Análise soja4

Soja 4	
Planta é ideal para as condições fornecidas?	Sim, a planta é ideal para as condições fornecidas
Taxa de sobrevivência	98,00%
Taxa de não sobrevivência	2,00%
Caso não sobreviva, foi encontrado soja similar na base de dados?	-----

Quadro 6: Análise soja5

Soja 5	
Planta é ideal para as condições fornecidas?	Não, a planta não é ideal para as condições fornecidas
Taxa de sobrevivência	25,00%
Taxa de não sobrevivência	75,00%
Caso não sobreviva, foi encontrado soja similar na base de dados?	Uma planta com maiores chances de sobrevivência é: Genetica1: 0.6 Genetica2: 0.6 Genetica3: 0.3 Genetica4: 0.5 Genetica5: 0.7

Quadro 7: Análise soja4

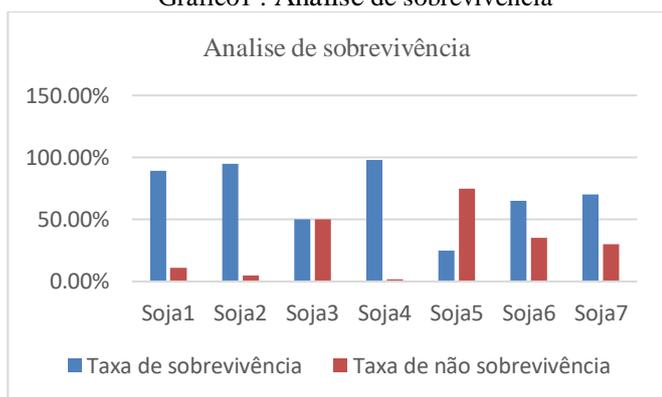
Soja 6	
Planta é ideal para as condições fornecidas?	Sim, a planta é ideal para as condições fornecidas
Taxa de sobrevivência	65,00%
Taxa de não sobrevivência	35,00%
Caso não sobreviva, foi encontrado soja similar na base de dados?	-----

Quadro 8: Análise soja7

Soja 7	
Planta é ideal para as condições fornecidas?	Sim, a planta é ideal para as condições fornecidas
Taxa de sobrevivência	70,00%
Taxa de não sobrevivência	30,00%
Caso não sobreviva, foi encontrado soja similar na base de dados?	-----

Com a análise dos quadros dos testes realizados, vimos que uma representação de cada situação que o sistema pode apresentar após avaliar os dados dos arquivos de características genéticas das novas sojas lidos. Menos não possuindo uma grande base de dados para a IA ser treinada analisando todos os cenários possíveis, podemos ver com o gráfico1, que das 7 sojas analisadas, 5 possuem probabilidades de sobrevivência, o que facilita o trabalho de pesquisadores para analisar as melhores sojas para certas condições climáticas.

Gráfico1 : Análise de sobrevivência



## V. Conclusão

Com o desenvolvimento desta pesquisa, podemos destacar a eficácia do modelo de Random Forest na classificação de plantas com base em características específicas. A utilização do Random Forest, demonstrou-se uma abordagem promissora para melhorar a precisão e a generalização do modelo. A análise detalhada das previsões, incluindo a avaliação da probabilidade de sobrevivência, forneceu percepções valiosas sobre o desempenho do modelo em diferentes cenários.

Além da capacidade preditiva, a abordagem adotada neste código não se limitou apenas à previsão binária, mas também incorporou uma análise mais abrangente, identificando plantas similares que sobreviveram e sugerindo alternativas quando necessário.

O sistema mostra uma ferramenta prática para orientar decisões relacionadas à sobrevivência de plantas em diferentes condições. O próximo passo agora é conseguir os dados genômicos das plantas no modelo FASTA e aprimorar o sistema para analisar a base de dados a partir desse modelo.

## REFERÊNCIAS

[1] MATIOLI, Sergio Russo; SOUZA, Diego Trindade de. **Introdução a Bioinformática**. Campinas: Unicamp, 2021. 175 p.

[2] PROSDOCIMI, Francisco; SANTOS, Fabricio R.. **Sobre bioinformática, genoma e ciência**. 2004. Ciência Hoje. Disponível em: <http://labs.icb.ufmg.br/lbem/pdf/prosdocimi04ch->

opinioao.pdf. Acesso em: 20 abr. 2023.

[3] ZERLOTINI NETO, Adhemar; NHANI JUNIOR, Antonio; VIEIRA, Fábio Danilo; CINTRA, Leandro Carrijo; MUDADU, Maurício de Alvarenga; FALCÃO, Paula Regina Kuser; GIACHETTO, Poliana Fernanda. **Aplicações da bioinformática na agricultura**. 2020. Disponível em: <https://www.alice.cnptia.embrapa.br/bitstream/doc/1126268/1/LV-Agricultura-digital-2020-cap10.pdf>. Acesso em: 15 abr. 2023.

[4] ARAËJO, Nilberto Dias de *et al.* **A ERA DA BIOINFORMÁTICA: SEU POTENCIAL E SUAS IMPLICAÇÕES PARA AS CIÊNCIAS DA SAÚDE**. 2008. Disponível em: <https://biblat.unam.mx/hevila/Estudiosdebiologia/2008/vol30/no70-72/16.pdf>. Acesso em: 10 mar. 2023

[5] MARTINS, Ana Caroline Pereira; ALVES, Joyce Micaelle; AGUIAR, Bruna Menezes; SOUZA, Andressa Samantha Oliveira. **A EVOLUÇÃO DA BIOINFORMÁTICA E SUA CONTRIBUIÇÃO PARA A SAÚDE**. 2015. Disponível em: [http://www.fepeg2015.unimontes.br/sites/default/files/resumos/arquivo\\_pdf\\_anais/a\\_evolucao\\_da\\_bioinformatica\\_e\\_sua\\_contribuicao\\_para\\_a\\_saude.pdf](http://www.fepeg2015.unimontes.br/sites/default/files/resumos/arquivo_pdf_anais/a_evolucao_da_bioinformatica_e_sua_contribuicao_para_a_saude.pdf). Acesso em: 29 mar. 2023.

[6] **O PROJETO GENOMA HUMANO**. [S. L.]: Revista Bioética, v. 5, n. 2, 04 nov. 2009. Trimestralmente. Disponível em: [https://revistabioetica.cfm.org.br/index.php/revista\\_bioetica/article/view/378/478](https://revistabioetica.cfm.org.br/index.php/revista_bioetica/article/view/378/478). Acesso em: 15 abr. 2023.

[7] **Compreendendo a digitalização do seu DNA**. 2014. Elaborado por Hospital Israelita Albert Einstein. Disponível em: [Compreendendo a digitalização do seu DNA](https://www.hisraelita.br/portal/verbo/verbo.php?verbo=1&id=1). Acesso em: 21 mar. 2023.

[8] **BIOINFORMATICS FOR HUMAN GENETICS: PROMISES AND CHALLENGES**. [S. L.], 8 fev. 2011. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1002/humu.21468>. Acesso em: 10 fev. 2023.

[9] V. A. . Bernardes, G. M. . Rodrigues, M. R. L. . Cabral, and L. L. F. G. . Neres, “The use of CRISPR-CAS9 Technique in Gene Therapy”, *RSD*, vol. 10, no. 14, p. e75101421778, Oct. 2021.

[10] SINGH, Vinay Kumar *et al.* **ROLE OF BIOINFORMATICS IN AGRICULTURE AND SUSTAINABLE DEVELOPMENT**. 2011. Disponível em: [https://www.researchgate.net/publication/240918582\\_ROLE\\_OF\\_BIOINFORMATICS\\_IN\\_AGRICULTURE\\_AND\\_SUSTAINABLE\\_DEVELOPMENT](https://www.researchgate.net/publication/240918582_ROLE_OF_BIOINFORMATICS_IN_AGRICULTURE_AND_SUSTAINABLE_DEVELOPMENT). Acesso em: 15 fev. 2023.

[11] GIACHETTO, Poliana Fernanda; HIGA, Roberto Hiroshi. **Bioinformática aplicada à agricultura**. 2014. Disponível em: <https://www.alice.cnptia.embrapa.br/bitstream/doc/1010705/1/capitulo0407614.pdf>. Acesso em: 7 jan. 2023.

[12] Esposito, A., Colantuono, C., Ruggieri, V. *et al.* Bioinformática para a agricultura na era do sequenciamento de próxima geração. *Chem. Biol. Technol. Agric.* **3**, 9 (2016). <https://doi.org/10.1186/s40538-016-0054-8>



**PUC  
GOIÁS**

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS  
GABINETE DO REITOR

Av. Universitária, 1069 • Setor Universitário  
Caixa Postal 86 • CEP 74605-010  
Goiânia • Goiás • Brasil  
Fone: (62) 3946.1000  
www.pucgoias.edu.br • reitoria@pucgoias.edu.br

**RESOLUÇÃO nº 038/2020 – CEPE**

**ANEXO I**

**APÊNDICE ao TCC**

**Termo de autorização de publicação de produção acadêmica**

O estudante GABRIEL HENRIQUE LIMA PICOUTO do  
Curso de ENGENHARIA DE COMPUTAÇÃO, matrícula 2019.10033.0035-7  
telefone: 64 98954 2870, e-mail GABRIELHENRIQUELIMAPICOUTO@GMAIL.COM,  
na qualidade de titular dos direitos autorais, em consonância com a Lei nº 9.610/98 (Lei  
dos Direitos do Autor), autoriza a Pontifícia Universidade Católica de Goiás (PUC Goiás)  
a disponibilizar o Trabalho de Conclusão de Curso intitulado  
BIOINFORMÁTICA E A UTILIZAÇÃO DA GENÉTICA NA AGRICULTURA  
\_\_\_\_\_, gratuitamente, sem ressarcimento dos direitos autorais, por 5 (cinco) anos,  
conforme permissões do documento, em meio eletrônico, na rede mundial de  
computadores, no formato especificado (Texto(PDF); Imagem (GIF ou JPEG); Som  
(WAVE, MPEG, AIFF, SND); Vídeo (MPEG, MWV, AVI, QT); outros, específicos da  
área; para fins de leitura e/ou impressão pela internet, a título de divulgação da produção  
científica gerada nos cursos de graduação da PUC Goiás.

Goiânia, 15 de DEZEMBRO de 2023.

Assinatura do autor: Gabriel Henrique Lima Picouto

Nome completo do autor: GABRIEL HENRIQUE LIMA PICOUTO

Assinatura do professor-orientador: Antônio Marcos de Melo Medeiros

Nome completo do professor-orientador: ANTÔNIO MARCOS DE MELO MEDEIROS