UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Lucas do Nascimento Macedo

Comparação de algoritmos de aprendizado de máquina para predizer futuras cepas do vírus da influenza

Uberlândia, Brasil

2023

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Lucas do Nascimento Macedo

Comparação de algoritmos de aprendizado de máquina para predizer futuras cepas do vírus da influenza

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Paulo Henrique Ribeiro Gabriel

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Ciência da Computação

Uberlândia, Brasil

2023

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Faculdade de Computação
Av. João Naves de Ávila, nº 2121, Bloco 1A - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4144 - http://www.portal.facom.ufu.br/ facom@ufu.br

ATA DE DEFESA - GRADUAÇÃO

Curso de Graduação em:	Bacharelado em Ciência da Computação				
Defesa de:	GBC082 - Projeto de Graduação 2				
Data:	22/06/2023	Hora de início:	14:00	Hora de encerramento:	15:05
Matrícula do Discente:	11811BCC044				
Nome do Discente:	Lucas do Nascimento Macedo				
Título do Trabalho:	Comparação de algoritmos de aprendizado de máquina para predizer futuras cepas do vírus da influenza				
A carga horária curricular foi cumprida integralmente?	(X) Sim ( ) Não				

Reuniu-se por meio da plataforma Microsoft Teams, da Universidade Federal de Uberlândia, a Banca Examinadora, designada pelo Colegiado do Curso de Graduação em Ciência da Computação, assim composta: Professores: Laurence Rodrigues do Amaral - FACOM/UFU; Maria Camila Nardini Barioni - FACOM/UFU; e Paulo Henrique Ribeiro Gabriel - FACOM/UFU orientador do candidato.

Iniciando os trabalhos, o presidente da mesa, Dr. Paulo Henrique Ribeiro Gabriel, apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu ao discente a palavra, para a exposição do seu trabalho. A duração da apresentação do discente e o tempo de arguição e resposta foram conforme as normas do curso.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

(X) Aprovado Nota: 95

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Laurence Rodrigues do Amaral**, **Professor(a) do Magistério Superior**, em 22/06/2023, às 15:04, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Maria Camila Nardini Barioni**, **Professor(a) do Magistério Superior**, em 22/06/2023, às 15:04, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Paulo Henrique Ribeiro Gabriel**, **Professor(a) do Magistério Superior**, em 22/06/2023, às 15:08, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4580155** e o código CRC **F50451A5**.

Dedico esse trabalho aos meus pais, Edna e Atenar, pois asfaltaram todo o caminho para que eu pudesse chegar até aqui e concluir meu curso.

Agradecimentos

Agradeço primeiramente aos meus pais, por serem responsáveis por tornar o sonho de cursar uma Universidade Pública possível, por me apoiarem em momentos difíceis que passei durante a vida acadêmica e por terem me mostrado desde cedo que a educação seria sempre meu passaporte para o futuro. Agradeço também a todos meus professores, sobretudo meu orientador, Prof. Paulo, por acreditar no meu potencial, me dar o privilégio de desenvolver um trabalho incrível e de um impacto social enorme como esse. Por fim, mas não menos importante, agradeço também aos meus amigos Miguel, Marcos e Gabriel por toda ajuda nos momentos de dificuldade, por cada conselho e por cada momento divertido que tornou essa caminhada mais leve.

"A matemática é a linguagem com a qual Deus escreveu o universo."

Galileu Galilei

Resumo

Todos os anos são feitos esforços para que a produção de vacinas que sejam eficazes contra o vírus influenza. No entanto, para que uma vacina tenha o efeito desejado, é necessário realizar estudos aprofundados sobre as possíveis mutações do vírus. Embora existam vários métodos de predição de novas cepas do vírus influenza na literatura, poucos ainda utilizam inteligência artificial como ferramenta nesse processo. Diante disso, este trabalho propõe a implementação de algoritmos de aprendizado para comparar seu desempenho na previsão de novas cepas, bem como compará-los com os modelos existentes na literatura. Os modelos de aprendizado foram implementados em *Python*, utilizando bibliotecas externas, como Florestas Aleatórias, Árvores Extras, *Naive Bayes* e Árvores de Decisão. Após a implementação, foram coletados dados das cepas H1N1 e H3N2 de um banco de dados e organizados em arquivos FASTA, seguindo uma ordem cronológica. A execução de todos os algoritmos revelou que o modelo de Florestas Aleatórias apresentou uma acurácia superior aos outros. Ao compará-lo com os modelos da literatura, constatou-se que o algoritmo de Florestas Aleatórias teve um bom desempenho, ficando atrás apenas de um modelo que utiliza aptidão preditiva.

Palavras-chave: Aprendizado de máquina, Predição, Influenza.

Abstract

Every year, efforts are made to produce vaccines that are effective against the influenza virus. However, in order for a vaccine to have the desired effect, in-depth studies on possible virus mutations are necessary. Although there are several methods in the literature for predicting new strains of the influenza virus, only a few utilize artificial intelligence as a tool in this process. In light of this, this study proposes the implementation of learning algorithms to compare their performance in predicting new strains and to compare them with existing models in the literature. The learning models were implemented in Python, using external libraries such as Random Forests, Extra Trees, Naive Bayes, and Decision Trees. After implementation, data from H1N1 and H3N2 strains were collected from a database and organized in FASTA files in chronological order. The execution of all algorithms revealed that the Random Forests model had a higher accuracy than the others. When compared to the models in the literature, it was found that the Random Forests algorithm performed well, second only to a model that utilizes predictive fitness.

Keywords: Machine learning, Prediction, Flu.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

A gripe ou influenza é uma doença respiratória causada por um vírus da família *Orthomyxoviridae* responsável por mais de 500 mil mortes por ano (SILVA FILHO et al., 2017; NOGUEIRA; PONCE, 2021). O nome "influenza" surgiu na Idade Média, pois se acreditava que alguns sintomas como febre, tosse e calafrio viriam de influencias planetárias (COMONE, 2011). Por volta de ano 495 a.C., Hipócrates relatou o primeiro caso de gripe que, durante algumas semanas, fez diversas vítimas e depois desapareceu (GRANATO; BELLEI, 2007). De acordo com Granato e Bellei (2007), estima-se que nos últimos 350 anos ocorreram nove pandemias causadas pelo vírus influenza no mundo, o que dá em média uma pandemia a cada 40 anos, aproximadamente. Uma das mais conhecidas foi relatada na Espanha em 1918 (GOMES, 2014), trazendo graves consequências à população. Conhecida como "Gripe Espanhola", essa pandemia coincidiu com o período marcado pela Primeira Guerra Mundial (1914–1918), o que fez com que os países negligenciassem o reconhecimento de um momento pandêmico, piorando, consequentemente, a gravidade da doença (SCHWARCZ; STARLING, 2020).

O vírus causador da Gripe Espanhola corresponde a uma variedade muito comum de vírus influenza: o H1N1. Sua descoberta foi possível a partir de cadáveres recuperados em regiões frias, o que possibilitou que fosse feito seu mapeamento molecular. Descobriu-se, ainda, que esse vírus tem origem aviária (GRANATO; BELLEI, 2007). Granato e Bellei (2007) mencionam que, cerca de 40 anos após a Gripe Espanhola, foram observadas mutações nesse vírus, o que ocasionou uma nova pandemia com uma nova cepa[1], a H2N2. Posteriormente, em meados do anos de 1967, outra cepa (H3N2) provocou uma nova pandemia.

De acordo com Saxena et al. (2012), os principais hospedeiros do vírus influenza são suínos e aves aquáticas. Esses animais atuam tanto como hospedeiros intermediários na transmissão inter-espécies como também contribuindo para o rearranjo genético do agente infeccioso. Alguns países do sudeste asiático possuem diversas espécies de aves aquáticas e suínos contamináveis por vírus aviário, permitindo com que a transmissão inter-espécies aconteça. Esses fatores torna um problema de saúde pois uma vez que um ser humano é infectado devido ao contato próximo a esses animais, o micro-organismo poderia sofrer mutações adicionais que o permita a transmissão inter-humana podendo assim provocar uma nova epidemia (GRANATO; BELLEI, 2007).

Todas as cepas de vírus influenza possuem estruturas proteicas diferentes. Dessas

[1] O termo "cepa" se refere a uma variante genética ou subtipo de um micro-organismos. Novas cepas podem surgir devido a mutações ou trocas de componentes genéticos quando dois ou mais vírus infectam uma mesma célula (YONG, 2013)

proteínas, existem duas muito importantes. A primeira é a hemaglutinina (HA) que é responsável por conectar o vírus à célula do hospedeiro. A hemaglutinina é bastante estudada pois tem uma importante função na patogenicidade e na resposta imunológica do indivíduo (COMONE, 2011). A segunda proteína é a neuraminidase (NA), responsável por permitir que o vírus penetre na célula do hospedeiro e se replique. Sua principal função é a disseminação da infecção pela degradação do ácido siálico e pela destruição de receptores de HA nas células hospedeiras, impedindo que a parte viral fique imobilizada na célula infectada (COMONE, 2011).

Todos os anos, estudos são feitos para tentar desenvolver uma vacina ou um medicamento que seja eficaz contra esse vírus. No entanto, essa classe de vírus está sempre sofrendo mutações, o que dificulta o desenvolvimento de um tratamento eficiente. Uma forma de prever quais as possíveis mutações que o vírus sofrerá é usando uma árvore filogenética. A partir de uma árvore filogenética podemos estudar toda a história evolutiva e as relações dos seres daquela árvore. Assim, partindo de um estudo filogenético do vírus influenza, é possível usar algoritmos que tentam predizer quais as próximas mutações baseadas em informações de ancestrais comuns.

Para manter a eficiência das vacinas, é necessário modelos de predileção efetivos. Segundo Koodli (2017) existem vários métodos que são usados para fazer predição, alguns deles são feitos usando fórmulas de predição que tem um pouco mais de sucesso. Porém, ainda segundo Koodli (2017), nenhum método utiliza a filogenética e o aprendizado de máquina juntos para tentar predizer as possíveis mutações das proteínas HA/NA. Diante disso, torna-se importante um estudo sobre o comportamento do vírus usando técnicas de aprendizado de máquina, tendo em vista que isso tem uma relevância social e científica.

Nesse contexto, este trabalho de conclusão de curso tem como objetivo aplicar técnicas de aprendizado de máquina para detecção de possíveis sequências das glicoproteínas hemaglutinina e neuraminidase presentes nas variantes H1N1 e H3N2 do vírus influenza. Assim, esse trabalho parte do pressuposto de que a inteligência artificial se apresenta como uma ferramenta consolidada capaz de contribuir com resultados bastante satisfatórios para este problema Para isso, inicialmente, foram selecionados dados do material genético do vírus influenza que serão utilizados durante o processo de testes. Essas informações foram obtidas no banco de dados *Influenza Research Database* (IRD)[2]. Em seguida, implementou-se quatro algoritmos de aprendizado de máquina. Considerou-se, nessa seleção, quatro algoritmos que demonstraram um bom desempenho em outros cenários (KOODLI, 2017): Árvores de Decisão, Florestas Aleatórias, Árvores Extras e *Naive Bayes*.

O passo seguinte e a implementação desses algoritmos e também sua execução. Esses algoritmos foram implementados em linguagem Python usando uma biblioteca vol-

[2] Disponível em: <https://www.fludb.org/>, acesso em 29 jun. 2022

tada para o aprendizado de máquina chamada *scikit-learn*[3]. Um vez implementado foram realizados os testes usando arquivos FASTA extraídos da IRD. Como resultado, observou-se que o algoritmo florestas aleatórias teve o melhor desempenho na predição das cepas H1N1 e H3N2.

Após a conclusão dos resultados, o algoritmo eleito como melhor para predição das cepas foi comparado com resultados obtidos na literatura. A comparação mostrou que o algoritmo florestas aleatórias tem um bom potencial em predições do vírus influenza, ocupando a segunda posição dos melhores métodos para previsão dos métodos comparados neste trabalho.

[3] Disponível em: <https://scikit-learn.org/stable/index.html>, acesso em jan. 2023

2 Referencial teórico

Neste capítulo, é apresentado o referencial teórico utilizado como base para o desenvolvimento deste trabalho, o mesmo se divide em conceitos básicos e trabalhos correlatos.

2.1 Inteligência artificial

A inteligência artificial (IA) é uma área relevante da ciência da computação, que oferece uma ampla gama de ferramentas para resolver problemas em diferentes setores. Entretanto, definir o conceito de IA pode ser desafiador devido à dificuldade em definir o significado do termo "inteligência" em si. Embora seja fácil identificar comportamentos inteligentes em humanos ou animais, isso se torna mais complexo quando se trata de programas de computador. Como afirmou o estudioso de Ciência da Computação Edgser Dijkstra, "perguntar se uma máquina pode pensar é o mesmo que perguntar se um submarino pode nadar", o que sugere que o objetivo final é o que importa e, se uma máquina alcança esse objetivo, ela pode ser considerada inteligente (FRANCO, 2014).

Para compreender melhor esse conceito, Fernandes (2004) fornece uma definição etimológica da IA. A palavra inteligência vem do latim *inter* e *legere*, que significam "entre" e "escolher", respectivamente. Portanto, pode-se concluir que a inteligência é a capacidade de um ser humano escolher entre uma coisa e outra. Já a palavra artificial também vem do latim *artificiale*, que significa algo que não é natural, ou seja, que foi criado ou construído por seres humanos. A partir dessa definição etimológica, é possível inferir que a inteligência artificial é a capacidade de um ser humano criar máquinas que possuam alguma habilidade que se assemelhe à inteligência humana.

Finalmente, de acordo com Russel e Norvig (2013), a IA é, em geral, um recurso capaz de automatizar e sistematizar tarefas complexas, o que a torna um instrumento valioso para qualquer esfera do conhecimento humano. A definição da IA pode ser dividida em duas dimensões principais: a primeira é composta por sistemas baseados no pensamento humano, e a segunda é composta por sistemas baseados na racionalidade.

Quanto às linhas de pesquisa em IA, diversos autores definiram quatro possíveis frentes de estudo. A primeira abordagem é a de um "sistema que pensa como ser humano", que envolve atividades que associamos ao pensamento humano, como a tomada de decisões, a resolução de problemas e o aprendizado. A segunda abordagem é a de um "sistema que atua como ser humano", que é a arte de criar máquinas que executam funções que exigem inteligência quando executadas por pessoas. A terceira abordagem é a de

um "sistema que pensa racionalmente", que é o estudo das faculdades mentais pelo uso de modelos computacionais. Por fim, a quarta abordagem é a de um "sistema que atua racionalmente", que é o estudo do projeto de agentes inteligentes (RUSSEL; NORVIG, 2013).

Desde o surgimento da IA, essas quatro linhas de pesquisa têm sido seguidas. Isso se deve às abordagens centradas nos seres humanos e às abordagens centradas na racionalidade. A abordagem centrada nos seres humanos é uma ciência empírica que envolve hipóteses e confirmação experimental, enquanto a abordagem centrada na racionalidade é empregada com conceitos matemáticos em sua construção (GOMES, 2014).

2.2 Aprendizado de máquina

Dentro da inteligência artificial encontra-se um subcampo denominado aprendizado de máquina. O aprendizado de máquina é definido como uma forma de predizer o futuro com base em acontecimentos do passado (DAUMÉ III, 2017). O aprendizado de máquina utiliza de técnicas para aprender e, esse aprendizado, acontece sempre que é aplicado uma mudança na estrutura do programa ou a base de dados é alterada, e o que promove essas mudanças são novos dados de entrada assim, isso pode garantir que o desempenho seja sempre aperfeiçoado.(NILSSON, 1996)

As técnicas de aprendizado se baseiam no método da indução, que é uma forma de inferência lógica para se obter conclusões a partir de um conjunto de exemplos (MONARD; BARANAUSKAS, 2003). O aprendizado indutivo é dividido em dois grupos: supervisionado e não-supervisionado. No aprendizado supervisionado o algoritmo de aprendizado recebe um conjunto de treinamento para que ele possa aprender e esse processo é chamado de treinamento. Posteriormente, o algoritmo é testado com outro conjunto de dados de teste onde é possível avaliar o seu desempenho (RUSSEL; NORVIG, 2013). Este método é o que será usado neste trabalho. Em um aprendizado não-supervisionado o algoritmo não passa por um treinamento, para aprender ele usa a entrada de dados e aprende a partir dela usando reconhecimento de padrões (RUSSEL; NORVIG, 2013).

Existem vários tipos de problemas de aprendizado, o que os diferenciam são os tipos de coisas que eles estão tentando prever. São classificados em: Regressão, Classificação binária, Classificação multi-classe e Ranking.(DAUMÉ III, 2017)

Regressão: Tenta prever um valor real, baseado em acontecimentos passados. Esse é o tipo de problema que proposto neste trabalho, isso porque baseado em amostras do material genético já existentes será possível prever uma nova ainda não conhecida.

Classificação binária: Tenta prever uma resposta binária *SIM* ou *NÃO*. Um exemplo disso é a caixa de e-mail que consegue determinar o que é spam e o que não é.

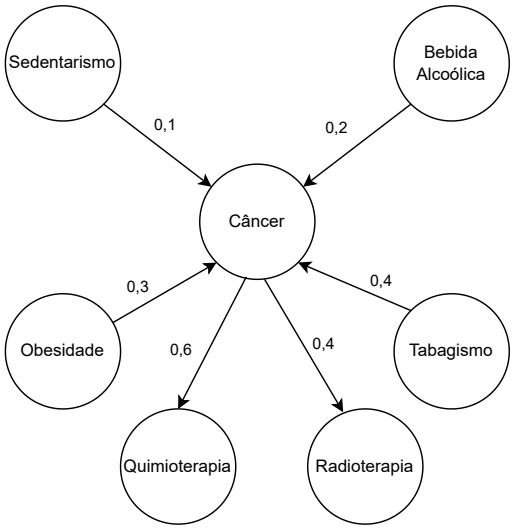
Classificação multi-classe: Tentar classificar um valor em um determinado conjunto de classes. Por exemplo, dado uma arara e um conjunto de classes contendo ave, mamífero, inseto. O algoritmo irá classificar arara como ave.

Ranking: Tentar ordenar um conjunto de dados por ordem de relevância. Por exemplo as matérias que um aluno precisa fazer para ir a um determinado evento.

2.2.1 *Naive Bayes*

O *Naive Bayes* é uma técnica de aprendizagem automática baseada em uma rede Bayesiana, que consiste em um modelo gráfico composto por um conjunto de probabilidades condicionais (JIANG; ZHANG; CAI, 2009). Essa representação é visualizada por meio de um grafo direcionado, no qual cada nó representa uma variável e cada aresta representa a relação probabilística entre as variáveis, como ilustrado na Figura 1.

Figura 1 – Um exemplo de um modelo de rede Bayesiana.



Fonte: Autoria própria

Na rede Bayesiana da Figura 1, cada variável é associada a uma probabilidade condicional, como a chance de um fumante ter câncer, que é de 40%. Já a chance de uma pessoa sedentária desenvolver câncer é de 10%, a de uma pessoa que faz uso de bebida alcoólica é de 20%, e a de uma pessoa obesa é de 30%. Além disso, uma pessoa com câncer tem 60% de chance de realizar quimioterapia e 40% de chance de realizar radioterapia. Com essas probabilidades condicionais, é possível calcular a probabilidade de uma pessoa fumante ter câncer, sabendo que ela realiza quimioterapia.

Para isso, utiliza-se o Teorema de Bayes, que é uma ferramenta que combina a fórmula de probabilidades totais com o teorema da probabilidade condicional (ARA-SOUZA, 2010). Esse teorema fornece uma previsão probabilística mais precisa, partindo de uma probabilidade *a priori*, como a eficácia de testes sorológicos em relação a um

agente patogênico desconhecido. O Teorema de Bayes é um corolário presente na lei de probabilidade total e pode ser expresso matematicamente pela Equação (2.1), em que $\Pr(A|B)$ é a probabilidade de A ocorrer sabendo que B ocorreu, $\Pr(B|A)$ é a probabilidade de B ocorrer sabendo que A ocorreu, $\Pr(A)$ é a probabilidade de A ocorrer e $\Pr(B)$ é a probabilidade de B ocorrer.

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)} \tag{2.1}$$

Com base na Equação (2.1), é possível responder à questão anterior. A resolução desse problema é apresentada na Equação (2.2), em que a probabilidade de uma pessoa fumante ter câncer, sabendo que ela realiza quimioterapia, é de 16%.

$$\Pr(A|B) = \frac{0{,}24 \times 0{,}4}{0{,}6} = \frac{0{,}096}{0{,}6} \Rightarrow \Pr(A|B) = 16\% \tag{2.2}$$

2.2.2 Árvores de decisão

Uma árvore de decisão é um modelo clássico de aprendizado de máquina que está relacionado ao conceito de dividir para conquistar (DAUMÉ III, 2017). Ela é definida usando uma estrutura em árvore, na qual cada nó intermediário representa uma pergunta (ou teste) e cada nó folha representa uma suposição. O funcionamento da árvore consiste em atribuir uma classe a um padrão que será recebido e filtrado nos nós intermediários por meio de perguntas. Para cada nó intermediário, a saída terá resultados mutuamente exclusivos e exaustivos (NILSSON, 1996).

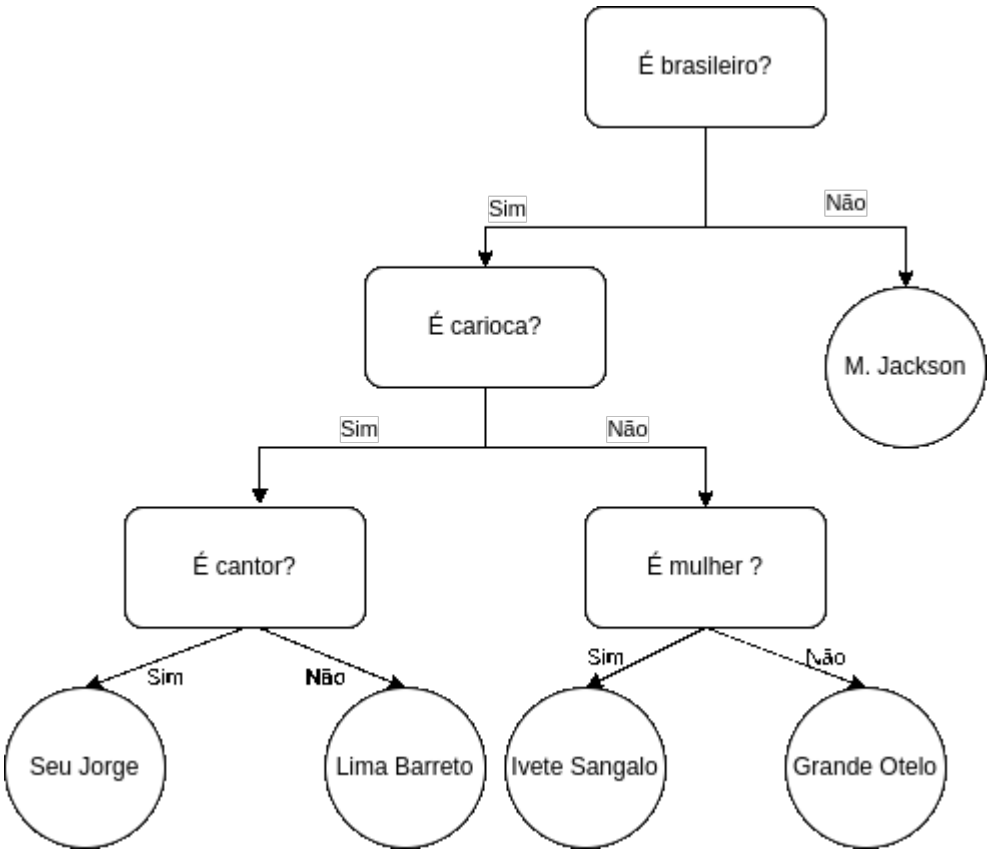
As árvores de decisão possuem algumas características, mas estas podem variar. A primeira delas é que o teste pode ser realizado de maneira multivariada, testando vários recursos da entrada de uma só vez, ou de maneira univariada, em que cada teste é feito usando um recurso da entrada por vez. Outra diferença é que, se para todos os testes de uma árvore de decisão for possível obter somente dois resultados, então essa árvore é chamada de árvore de decisão binária. Uma terceira característica é que os atributos de cada nó podem ser categóricos ou numéricos. Por fim, uma árvore de decisão pode ser chamada de árvore de decisão booleana se tiver duas classes e for uma árvore de decisão binária (NILSSON, 1996).

Na Figura 2, é apresentado um exemplo de árvore de decisão binária baseado no jogo *Akinator*[1]. Neste jogo, a máquina tenta descobrir o nome de uma pessoa em que o jogador está pensando. Para isso, a máquina faz várias perguntas ao jogador até chegar na pessoa que o jogador pensou. Observa-se que, na imagem, a máquina pergunta ao usuário se a pessoa é brasileira. Caso a resposta seja positiva, a máquina faz novos questionamentos tentando filtrar cada vez mais as perguntas até chegar nos nós folhas.

[1] Disponível em: <pt.akinator.com>. Acesso em 7 abr. 2023.

Caso a resposta do jogador seja negativa, a máquina chega a um nó folha concluindo que a pessoa que o jogador pensou é M. Jackson.

Figura 2 – Um exemplo de árvore de decisão binária.



Fonte: Autoria própria

2.2.3 Florestas aleatórias

Segundo Biau e Scornet (2016), floresta aleatória é uma técnica de aprendizado proposta por Breiman (2001) e consiste em um conjunto de classificadores estruturados em árvores, que pode ser descrito matematicamente pela Equação (2.3). Esta técnica tem um bom desempenho em modelos de classificação e regressão de propósito geral (BIAU; SCORNET, 2016).

$$\{h(x, \Theta_k), k = 1, \ldots\} \tag{2.3}$$

sendo que:

- Θ_k: São vetores aleatórios distribuídos de forma independente.
- x: É a entrada.

Em uma floresta aleatória para a k-ésima árvore é gerado um vetor aleatório Θ_k e este é um vetor independente, mas que possui a mesma distribuição dos vetores $\Theta_1,\Theta_2,\ldots,\Theta_{k-1}$. A partir deste passo a árvore é preenchida usando o conjunto de treinamento k, assim, quando a árvore recebe uma entrada x é possível aplicar a Equação (2.3) resultando em um classificador (BREIMAN, 2001).

O algoritmo floresta aleatória descrito por Daumé III (2017), inicia construindo várias árvores de classificação. O conjunto de treinamento são vetores que são usados nas ramificações de cada árvore e selecionados aleatoriamente por meio de substituição, isso significa que em algumas ramificações possa ter o mesmo vetor (DAUMÉ III, 2017). Segundo Morais (2017), isso revela que existe uma probabilidade que a partir de n seleções independentes, com reposição, um indivíduo pode não ser selecionado, e essa probabilidade é dada pela equação 2.4.

$$\prod_{i=1}^{n} \frac{n-1}{n} = 1 - \frac{1}{n}^{n} \rightarrow e^{-1} \approx 0{,}367 \tag{2.4}$$

Note que um pouco mais de um terço dos dados não são selecionados. Esse conjunto de dados não selecionados recebe o nome de *out-of-bag* (OOB) e esses são muito importantes para tentar estimar o erro do modelo de predição (MORAIS, 2017). De acordo com Morais (2017), em aprendizado de máquina o conjunto de OOB é denominado de conjunto de validação.

Ainda durante o treinamento, os nós folhas dessas árvores são preenchidos com base no dados que foram aprendidos. Dessa forma, em concordância com Daumé III (2017), o resultado da classificação é uma votação de todas as árvores. Para o algoritmo é necessário apenas três argumentos. os dados, a profundidade da árvore e o número de árvores. No Algoritmo 1 é apresentado o pseudocódigo desse modelo de classificação.

Algoritmo 1: Pseudocódigo do algoritmo de floresta aleatória

Função `FlorestaAleatoria`(S, F, B)**:**

$H \longleftarrow \emptyset$

para $i \in 1 \ldots B$ **faça**

$S^{(i)} \leftarrow$ Um conjunto de treinamento de S

$h_i \leftarrow$ `TreinaArvoreAleatoria`(S^i, F)

$H \leftarrow H \cup h_i$

fim

retorna H

Função `TreinaArvoreAleatoria`(S, F)**:**

para cada *nó* $\in S^{(i)}$ **faça**

$f \leftarrow$ Um pequeno subconjunto de F

$T \leftarrow$ Divide o melhor subconjunto de F

fim

retorna *Árvore treinada*

2.2.4 Árvores Extras

As Árvores Extras são uma técnica de aprendizado de máquina supervisionado baseada em conjuntos de árvores de decisão ou regressão. Diferentemente de outros algoritmos baseados em árvore, as Árvores Extras dividem os nós selecionando pontos de corte totalmente aleatórios e utilizam todo o conjunto de treinamento para aumentar as árvores (GEURTS; ERNST; WEHENKEL, 2006). Elas foram desenvolvidas como uma ampliação do algoritmo Florestas Aleatórias (GEURTS; ERNST; WEHENKEL, 2006). Assim, a técnica de Árvores Extras utiliza o mesmo preceito das Florestas Aleatórias e, para treinar cada estimador base, ela utiliza um subconjunto aleatório. Porém, para selecionar a melhor característica, ela escolhe junto com o valor correspondente à divisão do nó (AHMAD; REYNOLDS; REZGUI, 2018). O Algoritmo 2, descrito por Geurts, Ernst e Wehenkel (2006), apresenta a implementação das Árvores Extras.

Ao comparar as Árvores Extras com outros algoritmos baseados em árvore que utilizam o método de aprendizagem conjunta, podemos notar que ela se diferencia pois divide os nós selecionando pontos de corte aleatórios e também porque utiliza toda a amostra para aprendizado a fim de aumentar a quantidade de árvores construídas. Esse algoritmo inicia recebendo dois parâmetros: o primeiro é o número de atributos selecionados aleatoriamente e o segundo é o tamanho mínimo da amostra para dividir o nó. Logo em seguida, o algoritmo cria um conjunto de árvores de decisão por meio da abordagem *top-down*, ou seja, a árvore é construída do nó raiz para os nós folhas.

Algoritmo 2: Pseudocódigo do algoritmo de Árvores Extras.

Função `ConstroiArvoreConjunto`(S)**:**
  Entrada: Conjunto de treinamento S
  Saída: Conjunto de árvores
  para $i \leftarrow 1$ **até** M **faça**
    $\tau_i \leftarrow$ `ConstruirArvoreExtra` (S)
  fim
  retorna τ

Função `ConstruirArvoreExtra`(S)**:**
  Entrada: Conjunto de treinamento S
  Saída: Uma árvore t
  se $|S| < n_{\min}$ ***ou*** *todos os atributos candidatos são constantes em* S ***ou*** *a variável de saída é constante em* S **então**
    retorna *Nó folha rotulado pela sua classe*
  fim
  senão
    $A_k \leftarrow S_k$
    para $i \leftarrow 0$ **até** K **faça**
      $S_i \leftarrow$ `PegarDivisaoAleatoria` (S, A_i)
    fim
    $S_l, S_r \leftarrow$ `DivideConjunto` (S)
    $t_l \leftarrow$ `ConstroiArvoreConjunto` (S_t)
    $t_r \leftarrow$ `ConstroiArvoreConjunto` (S_r)
    $S_* \leftarrow$ Nova árvore com t_l à esquerda e t_r à direita
  fim
  retorna *Árvore* S_*

A função *ConstruirArvoreExtra* é executada várias vezes utilizando a amostra de dados completa para que seja possível gerar um modelo de conjunto. A partir disso, as árvores são unificadas para produzir uma previsão final. Em problemas de classificação, o resultado é produzido a partir da maioria de votos em comum de todas as árvores. Já em problemas de regressão, o resultado é obtido através da média aritmética dos resultados de todas as árvores.

Geurts, Ernst e Wehenkel (2006) analisam esse pseudo-código a partir de duas perspectivas: viés-variância e computacional. De acordo com eles, a randomização no ponto de corte combinada com a média do conjunto pode ser capaz de reduzir a variância em comparação a outros métodos de randomização aplicados por outros meios. Em uma perspectiva computacional, o crescimento das árvores tem uma complexidade $n \log n$ em comparação ao tamanho da amostra de aprendizado. Contudo, Geurts, Ernst e Wehenkel (2006) mencionam ainda que o procedimento de divisão de nós é simples e que isso faz com que o fator constante seja menor em comparação a outros métodos baseados em conjunto de amostras.

2.3 Vírus

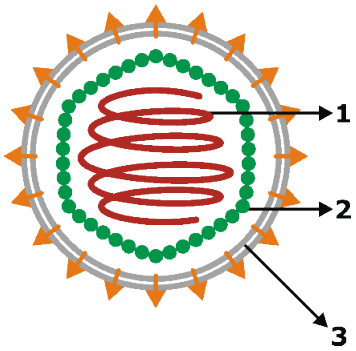
Um vírus é um organismo acelular, ou seja, que não possui célula, e atua como um parasita intracelular no organismo do hospedeiro. Eles são capazes de infectar animais, plantas, fungos, bactérias, arqueias e seres humanos (CARTER; SAUNDERS, 2013). Os vírus são agentes patogênicos de várias doenças, como, por exemplo, influenza, raiva, AIDS, entre outras (BRASIL, 2010).

Segundo Carter e Saunders (2013), a primeira indicação da existência desses agentes infecciosos foi feita no Século XIX por dois cientistas: Martinus Beijerinck e Dimitri Ivanovski. Essa descoberta foi feita analisando uma planta que estava "doente". Assim, eles retiraram extratos dessa planta e os passaram por filtros. Logo em seguida, foi observado que o material filtrado continha um agente capaz de infectar plantas saudáveis. Os cientistas eliminaram qualquer possibilidade do agente infeccioso ser uma bactéria, pois depois de filtrado, não é possível cultivar bactérias. Foi descartada também a possibilidade de ser uma toxina, pois o agente infeccioso permanecia prejudicial após várias transmissões para plantas saudáveis. Foi então que a partir dessa observação, Beijerinck deu o nome vírus ao agente infeccioso.

Alguns anos depois, Freidrich Loeffler e Paul Frosch fizeram outro experimento no qual conseguiram transmitir febre aftosa (*Aphtovirus*) de um animal para o outro. Mais tarde, Walter Reed e James Carroll mostraram que o agente causador da febre amarela é um agente filtrável (CARTER; SAUNDERS, 2013). Isso explicaria a análise feita por Martinus Beijerinck e Dimitri Ivanovski, quando notaram que após filtrar o extrato da planta, os agentes ainda estavam presentes, reforçando que estavam lidando com um agente viral.

Os vírus são estruturas compostas por um genoma e uma cápsula proteica, conforme ilustrado na Figura 3. O genoma é formado por ácidos nucleicos, podendo ser de DNA ou RNA. O DNA é responsável pelo armazenamento das informações genéticas, enquanto o RNA é responsável pela produção de proteínas. Os nucleotídeos são as unidades básicas dessas macromoléculas, compostos por uma pentose, um fosfato e uma base nitrogenada. A diferença fundamental entre DNA e RNA está na estrutura da pentose, sendo que a do DNA é formada por desoxirribose e a do RNA é formada por ribose. Por sua vez, a cápsula proteica dos vírus é composta por lipídeos, moléculas de gordura orgânica insolúveis em água. Alguns vírus possuem, ainda, um envelope viral, ou seja, uma estrutura presente na parte externa à cápsula proteica. É nessa estrutura que são encontradas as glicoproteínas de interesse deste trabalho, ou seja, a hemaglutinina (HA) e a neuraminidase (NA).

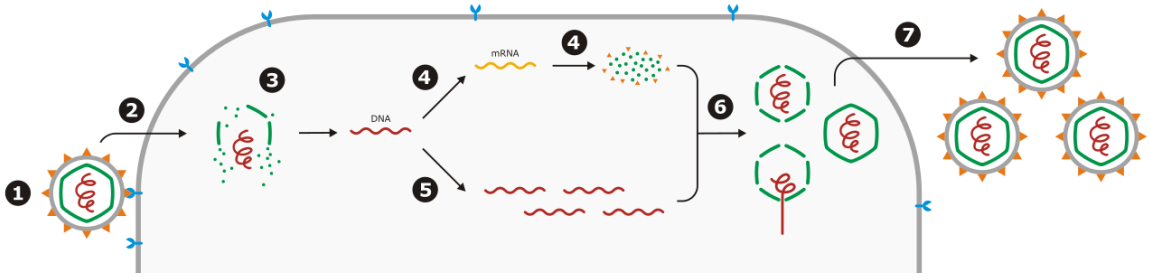
Figura 3 – Representação da estrutura de um vírus. Nessa figura, 1 indica o ácido nucleico (DNA ou RNA), 2 é a cápsula proteica e 3 é o envelope viral.



Fonte: Adaptado de Nossedotti (2011b).

O processo de replicação dos vírus envolve diversas etapas, iniciando com a adsorção, que consiste na ligação do vírus a receptores primários e secundários da célula hospedeira. A seguir, ocorre a fusão do envoltório do vírus com a membrana citoplasmática da célula hospedeira, permitindo a entrada do vírus na célula. Na terceira etapa, ocorre o desnudamento, ou seja, a liberação do material genético do vírus dentro da célula. Em seguida, há a transcrição do genoma viral em DNA ou RNA, seguida da transcrição do mRNA, que é responsável por produzir as proteínas virais. A tradução dessas proteínas é a penúltima etapa, seguida da montagem e liberação dos vírus, que podem deixar a célula hospedeira por meio do brotamento. A Figura 4 ilustra esse processo.

Figura 4 – Ilustração do processo de replicação de um vírus dentro da célula hospedeira.



Fonte: Adaptado de Nossedotti (2011a).

Vale destacar que existem diferentes tipos de vírus, que podem infectar diferentes tipos de células e apresentar diferentes mecanismos de replicação. Alguns vírus podem integrar seu material genético ao genoma da célula hospedeira e permanecer latentes por longos períodos de tempo, enquanto outros podem induzir a célula hospedeira a produzir novas partículas virais sem matá-la imediatamente.

Um dos principais desafios na luta contra as doenças virais é a capacidade dos vírus de se adaptar e evoluir rapidamente. O vírus influenza, por exemplo, apresenta constantes

mutações em seu material genético, dificultando a criação de vacinas eficazes contra todas as suas variantes. Por isso, é essencial que haja uma constante pesquisa e desenvolvimento de novas estratégias de prevenção e tratamento de doenças virais.

2.4 Trabalhos correlatos

Um estudo feito por Koodli (2017) analisou o comportamento das cepas H1N1 e H3N2 do vírus influenza usando aprendizado de máquina. Para isso, foram implementados algoritmos que observam como as glicoproteínas hemaglutinina (HA) e neuraminidase (NA) evoluíram de uma geração para outra, e partir disso foi construído uma árvore de decisão que ajuda a prever uma nova sequência de HA e NA. Foram implementados três algoritmos de aprendizado, são eles: árvores de decisão, florestas aleatórias e árvores extras. Os resultados mostraram que o algoritmo florestas aleatórias teve um melhor desempenho quando comparando a árvore de decisão e árvores extras que, por sua vez, tiveram resultados muito semelhantes. Árvores de decisão teve uma acurácia média 56% para H1N1 e 70% para H3N2. Já o algoritmo florestas aleatórias teve uma acurácia média de 84% para H1N1 e 90% para H3N2. Isso mostra que o algoritmo florestas aleatórias teve a melhor precisão quando comparado aos outros e também que algoritmos de aprendizado podem ter um bom desempenho na previsão de futuras cepas da gripe.

Outro estudo feito por Łuksza e Lässig (2014) desenvolveu um modelo de aptidão preditiva para hemaglutinina, onde a partir dele é possível prever a evolução de uma população viral anualmente. Para determinar a aptidão da cepa, dois fatores são muito importantes: mudanças adaptativas de epítopos e mutações deletérias fora dos epítopos. A partir desses fatores é possível determinar a aptidão para as cepas que estão presentes em uma determinada época usando os dados genéticos de todas as cepas anteriores. Com base na aptidão e frequência de cada linhagem, é possível prever a frequência das linhagem descendentes. Portanto esse modelo faz o mapeamento o histórico adaptativo da influenza A. Os resultados se mostraram bastante satisfatórios sendo 63% e 93% para as cepas H1N1 e H3N2, respectivamente.

Um estudo desenvolvido por Neher et al. (2016) usa propriedades antigênicas do vírus influenza medidas em ensaios de inibição de hemaglutinação(HI), assim é possível obter a concentração mínima de antissoro para impedir a reticulação de glóbulos vermelhos pelo vírus com base em unidades hemaglutinantes. Para a predição de títulos HI foram utilizados dois métodos, o primeiro é o modelo de árvore que mostra títulos de HI como uma soma de contribuições associadas com ramificações internas na árvore filogenética que conectam o vírus de referência com o vírus de teste. O segundo modelo é chamado de modelo de substituição, ele apresenta títulos de HI em termos de uma soma de contribuições associadas a substituições de aminoácidos entre o vírus de referência e o

vírus de teste. Esses modelos conseguiram prever as cepas H1N1 e H3N2. Para H1N1 foi obtido uma precisão de 77%, já para a cepa H3N2 foi obtido uma precisão de 72%.

Bush et al. (1999) desenvolveram um método de predição da evolução do vírus influenza. Para isso, eles usaram um modelo de predição que usa dados genéticos sem a necessidade das informações sobre as propriedades antigênicas do vírus. Os dados genéticos para esse estudo foram coletados em meados do final década de 80 e início da década de 90. Esse método utiliza uma árvore filogenética construída usando genes da hemaglutinina do subtipo H3 do vírus influenza A e analisa ao longo do tempo onde existe uma "linhagem de tronco", i. e., uma linhagem que todas as cepas subsequentes surgem dela. Ele ainda estipula que a linhagem que mais sofrer substituições de aminoácidos nos 18 códons ao longo de toda a ramificação será a linhagem de tronco. Porém, para verificar se a previsão está correta, Bush et al. (1999) observaram durante 11 períodos de gripe o comportamento desse método para predição de novas linhagem. A partir disso, foi possível predizer a linhagem de 9 períodos dos 11 analisados. Segundo os autores, os outros períodos onde não houve uma predição correta deve-se ao fato de que o método escolheu o mesmo isolador preditivo. Sugerem ainda que, isso poderia ser corrigido desenvolvendo critérios para linhagens que se tornarem extintas. Diante disso, foi possível dizer que houve uma taxa de predição correta de 81%.

Quando Bush et al. (1999) desenvolveram o método preditivo, notaram que usar o mesmo isolador preditivo diminui a precisão de método. Na tentativa de solucionar esse problema, Suzuki (2013) desenvolveu um método teórico que é capaz prever a evolução antigênica da subvariante H3N2 do vírus influenza A, avaliando mutações através de estimativas de distância antigênica. Essa diferença antigênica entre as cepas pode ser definida usando títulos de inibição de hemaglutinação. Na Equação (2.5) é apresentado o cálculo que nos permite determinar a distância antigênica entre cepas (LEES; MOSS; SHEPHERD, 2010)

$$d_{ij} = \frac{1}{2} \log \left(\frac{t_{ij}}{t_{ii}} \frac{ji}{jj} \right) \tag{2.5}$$

sendo que:

- i e j: São linhagens do vírus influenza A.
- t: É o limiar de escape antigênico.

No desenvolvimento do seu método, Suzuki (2013) dividiu os dados em grupos de treinamento e validação. Para o treinamento, os valores foram otimizados reduzindo a soma dos mínimos quadrados na estimativa da distância antigênica de todos os pares disponíveis das cepas i e j e, logo em seguida, fazendo uso de algoritmos genéticos. Nos testes foi usado o melhor modelo antigênico obtido no treinamento para prever sequência

de aminoácidos de hemaglutinina tipo 1 para as cepas predominantes no ano de 2001. Como resultado, Suzuki (2013) notou que o desempenho de um dos modelos, o qual nomeou como modelo 5, pareceu ser mais próximos de estudos feitos anteriormente. Nesse modelo, as características principais que auxiliam na melhor predição dos aminoácidos é a sensibilidade e a especificidade. Essas características tiveram os seguintes resultados: sensibilidade baixa, aproximadamente 0,593 e a especificidade alta próximo a 0,9. Logo, é possível concluir que esse modelo pode complementar na identificação de variantes antigênicas (SUZUKI, 2013).

Usando um gráfico dinâmico por varredura, ou do inglês, *sweep dynamics plot* (SD) Klingen et al. (2018) desenvolveram um método para monitorar a adaptação do vírus influenza. Um SD combina algoritmos filogenéticos e ferramentas estatísticas para descrever como ocorre a alteração molecular de um agente infeccioso durante um período usando dados de sequência longitudinal. Nesse estudo, foi investigado a adaptação genômica do vírus influenza A dos subtipos H1N1 e H3N2 da pandemia de 2009. Analisando a cepa H1N1, foi examinado algumas sequências de aminoácidos, nucleotídeos e dez proteínas coletadas de 2009 a 2015.

De acordo com Klingen et al. (2018), a análise indicou que nas proteínas de superfície NA e HA foi encontrado as maiores varreduras seletivas, enquanto que na proteína NS2 foi encontrado a menor quantidade. Uma varredura seletiva é um processo onde uma mutação benéfica passa a ser mais frequente na população levando a redução ou eliminação da variação genética (PENNINGS; HERMISSON, 2006). Foi notado também que existia fortes ligações entre os segmentos 32-34 o que resulta em mudanças neutras que são incluídas em outros segmentos. Com os gráficos SD foi possível identificar essas mudanças neutras e determinar se o processo foi causado por varredura. Dessa forma, os gráficos SD constataram que alterações que se agrupavam na estrutura da proteína juntamente com sítios, i. é. uma região da enzima onde as moléculas do substrato se ligam para sofrer uma reação química, apontavam uma potencial relevância no estudo desse processo.

Ainda segundo Klingen et al. (2018), o estudo apresentou informações e conclusões sobre a cepa H3N2. Para esta cepa, foram coletados sequências de 34 amostras durante os anos de 1999 a 2015. A partir dessas amostras foi possível observar que em várias amostras, havia alterações relacionadas à varredura seletiva na proteína HA. Com esses registros, foi investigado se para cada varredura seletiva encontrada indicava o surgimento de uma cepa antigenicamente nova, i. e., cepas que que poderiam escapar de agentes antivirais. Foi apurado ainda a coerência entre as varreduras detectadas e as novas variantes antigênicas, comparando sítios à varredura com sítios de mudanças de antigenicidade. Por fim, depois de todas as analises foi concluído que a maioria das alterações relacionadas a varredura seletiva que foi identificada na proteína HA ocorreu em locais conhecidos de evasão imune.

Com o uso de uma rede de transição de sítios, do inglês, *site transition network*

(STN) baseado em informações mútuas, Xia et al. (2009) mapearam a evolução genética da variante H3N2 do vírus influenza. Nesse estudo os dados coletados foram todas as sequencias HA tipo 1 da variante H3N2 até o ano de 2008. Ao final do processo de coleta e separação dos dados, foram coletados 4064 sequências. A partir desse passo foi possível calcular o método de informação mútua. O método de informação mútua, do inglês, *mutual information* (MI), é um algoritmo de inferência de rede que é utilizado para calcular a correlação entre dois locais de resíduos, a Equação (2.6) apresenta como o cálculo do MI pode ser feito. A partir dele é factível projetar uma STN.

$$I(x, y) = S(x) + S(y) - S(x, y) \tag{2.6}$$

sendo que:

- x e y: Variáveis discretas que representam a mutação de sítios.
- $S(t)$: É a entropia de uma variável arbitrária t.

Após projetado a rede STN, foi possível notar que grande parte das interações dinâmicas estão próximas dos epítopos e regiões de domínio de ligação ao receptor. É possível verificar também que as alterações antigênicas acumulam durante o tempo e isso se deve por conta de várias mutações que aconteceram em locais antigênicos. Para além disso, a pesquisa revelou que após feito a subdivisão do STN em várias sub-redes, foi possível ter um olhar mais detalhado sobre as características da mudança antigênica. Em conclusão, utilizando STNs Xia et al. (2009) conseguiu uma precisão de 70% quando analisou as mutações da cepa H3N2 nos anos entre 2003 a 2004. Além conseguir prever sete cepas futuras para o períodos de 2009 a 2010.

3 Desenvolvimento

O desenvolvimento deste trabalho teve início com a implementação dos algoritmos de aprendizado de máquina propostos por Koodli (2017), os quais serviram como base para o desenvolvimento desta pesquisa. A implementação foi realizada em *Python*, utilizando os recursos da biblioteca *scikit-learn* (PEDREGOSA et al., 2011), e envolveu os seguintes algoritmos: florestas aleatórias, árvores extras, árvores de decisão e *Naive Bayes*. Durante a implementação, observou-se que todos esses algoritmos requerem valores inteiros como entrada, o que demandou a criação de uma classe para a conversão dos dados de entrada, que estão no formato de sequências de material genético representadas por *strings*, em valores inteiros.

Como entrada para esses algoritmos, foram utilizados arquivos do tipo FASTA, os quais contêm o material genético de um conjunto de cepas[1]. Esses arquivos foram obtidos no *Influenza Research Database* (IRD), um banco de dados disponível publicamente para pesquisar, analisar, visualizar, salvar e compartilhar dados para pesquisa de vírus influenza[2]. A seleção das cepas foi realizada utilizando o critério tempo, ou seja, optando pelas cepas mais recentes.

O material genético é composto por uma sequência de nucleotídeos representados pelas letras iniciais das bases nitrogenadas: guanina (G), citosina (C), adenina (A) e timina (T). Desse modo, cada arquivo FASTA é composto por uma sequência desses caracteres, além de um cabeçalho, também textual, conforme ilustrado na Tabela 1. No entanto, os algoritmos de aprendizado de máquina requerem conjuntos de treinamento e teste com valores numéricos reais, gerando um erro caso o tipo de dado não seja compatível com o esperado. Portanto, uma classe de conversão foi criada com o objetivo de transformar os dados de entrada em valores numéricos reais, atribuindo a cada base nitrogenada um número inteiro específico. Dessa forma, a função de conversão pode retornar a sequência no formato aceito pelos algoritmos de aprendizado. A Figura 5 ilustra o processo de conversão descrito.

[1] FASTA é um formato de arquivo de texto para armazenar dados de sequência macromolecular (GIBAS; JAMBECK, 2001)

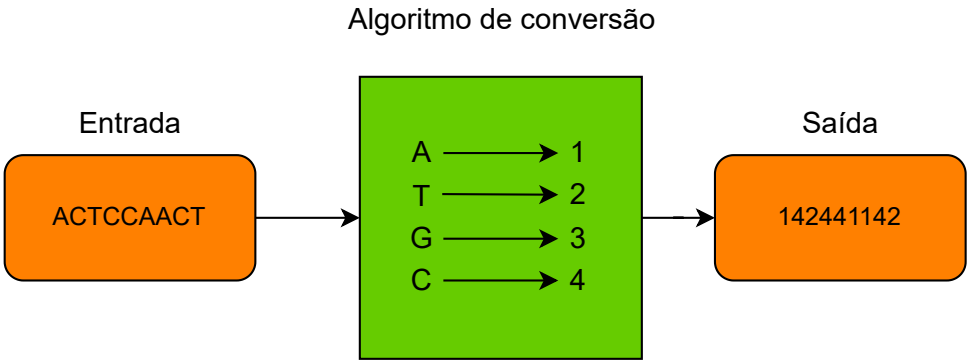
[2] O IRD é financiado pelo Instituto Nacional de Alergia e Doenças Infecciosas (NIAID), vinculado ao Instituto Nacional de Saúde (NIH), dos Estados Unidos.

Tabela 1 – Exemplo de arquivo FASTA. Esse arquivo se refere às primeiras três linhas do genoma do vírus H1N1, em uma cepa sequenciada em Tottori, Japão.

```
>accn|LC638170 Influenza A virus (H1N1) A/Tottori/ST3310/2020 HA
gene for hemagglutinin, complete cds. [Influenza A virus (H1N1)
A/Tottori/ST3310/2020 | 1323429.594]
atgaagacaatactagtagttctgctgtatacatttacaaccgcaaatgcagacacatta
tgtataggttatcatgcgaacaattcaacagacactgtagacacagtactagaaaagaat
gtaacagtaacacactctgtcaatcttctggaagacaagcataacggaaaactatgcaaa
```

Fonte: Obtido no *Influenza Research Database*.

Figura 5 – Ilustração do processo de conversão dos dados de sequência de material genético para valores numéricos reais utilizando o conversor implementado.



Fonte: Autoria própria

Para o desenvolvimento dos modelos de predição, definiu-se que os dados seriam divididos em 70% para treinamento e 30% para teste. Além disso, foi estabelecido que todos os modelos de predição deveriam utilizar o coeficiente de determinação R^2 como métrica de avaliação, pois essa métrica permite avaliar a precisão do modelo e ajustar os dados para obter uma maior acurácia (MARTINS, 2018). Com essas definições, deu-se início à implementação dos algoritmos de florestas aleatórias, árvores de decisão e árvores extras utilizando o modelo de regressão, uma vez que esse modelo apresenta bom desempenho para prever valores contínuos. O algoritmo *Naive Bayes* foi implementado utilizando o modelo Gaussiano, pois a biblioteca utilizada para este método não disponibiliza o modelo de regressão e *Naive Bayes* é um classificador, ou seja, o resultado é binário (0 ou 1) e não contínuo (por exemplo, 0.2). Em seu estudo, Frank et al. (2000) mostrou que o uso do algoritmo *Naive Bayes* com o modelo de regressão não produz resultados satisfatórios. Portanto, mesmo que fosse implementado o algoritmo utilizando regressão linear ou regressão logística, os resultados seriam inadequados.

Após a implementação desses algoritmos, foi realizada a seleção do material genético do vírus influenza. Foi estabelecido que a quantidade preferencial de amostras para

treinamento e teste seria de 50 amostras para cada cepa, pois esse número permitiria um bom desempenho de predição. Ao analisar os dados disponíveis nas bases de dados, constatou-se que havia uma grande quantidade de amostras da cepa H3N2, porém uma quantidade muito baixa de amostras da cepa H1N1. De acordo com o Centro de Controle e Prevenção de Doenças (CDC, do inglês *Centers for Disease Control*) dos Estados Unidos[3], desde outubro de 2022, aproximadamente 77% de todas as amostras recebidas eram da cepa H3N2. Isso indica que a cepa dominante atualmente é a H3N2 e justifica a escassez de amostras da cepa H1N1.

Dessa forma, para a cepa H3N2, foram coletadas duas proteínas de superfície (hemaglutinina e neuraminidase) nos anos de 2021 a 2022. Após a coleta, as proteínas foram agrupadas e, em seguida, foram criados arquivos de treinamento e teste para cada uma das estruturas do vírus. Ao final, obteve-se quatro arquivos FASTA, dois para hemaglutinina e dois para neuraminidase, referentes aos conjuntos de treinamento e teste de cada uma delas. A Tabela 2 apresenta algumas cepas que fazem parte do experimento, e a partir dessa tabela, podem ser observadas informações relevantes, como o local e o ano de coleta

Tabela 2 – Amostras usadas para o experimento com a cepa H3N2.

Local	Ano	Segmento	Dados
Flórida	2021	H3N2-HA	Treinamento
Califórnia	2022	H3N2-HA	Teste
Pensilvânia	2021	H3N2-NA	Treinamento
Califórnia	2022	H3N2-NA	Teste

Fonte: Autoria própria.

Considerando a baixa quantidade de cepas de H1N1 disponíveis na base de dados, foi necessário utilizar uma quantidade diferente de amostras em comparação com a cepa H3N2. Para a hemaglutinina da cepa H1N1, foram utilizadas 24 amostras para treinamento e 24 amostras para teste. As amostras de treinamento foram coletadas nos anos de 2019 e 2020, enquanto as amostras de teste foram coletadas no ano de 2021. Para a proteína neuraminidase, também foram coletadas 24 amostras para treinamento e 24 amostras para teste. As amostras de treinamento foram coletadas no ano de 2021 e as amostras de teste foram coletadas no ano de 2022. Dessa forma, foram obtidos quatro conjuntos de amostras da mesma cepa, divididos em dois grupos de proteínas diferentes. A Tabela 3 apresenta algumas cepas que fazem parte dos experimentos descritos nesta monografia.

Com os dados selecionados e os algoritmos implementados, o programa para predição dessas proteínas foi executado para cada grupo de treinamento e teste. A Figura 6

[3] Disponível em: <https://www.cdc.gov/>, acesso em 10 jan. 2023

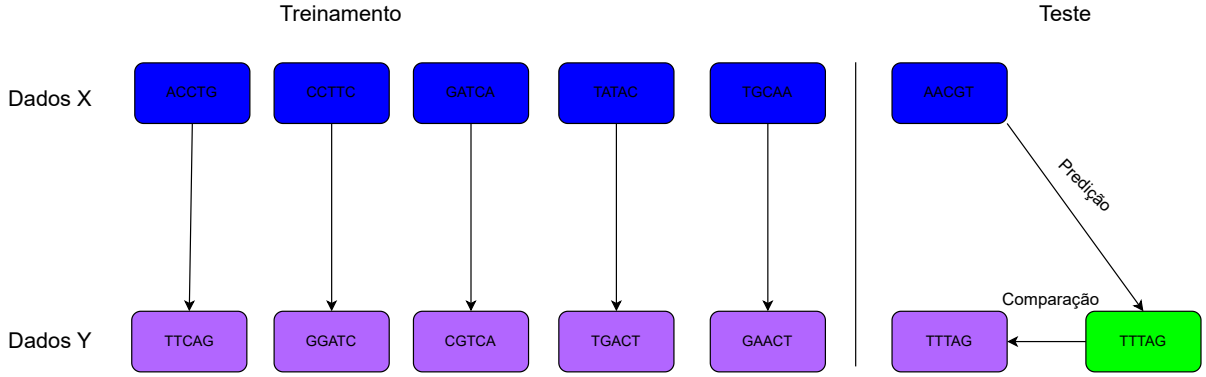
Tabela 3 – Amostras usadas para o experimento com a cepa H1N1.

Local	Ano	Segmento	Dados
Tottori	2020	H1N1-HA	Treinamento
Coreia do Sul	2021	H1N1-HA	Teste
Wenzhou	2019	H1N1-NA	Treinamento
Michigan	2021	H1N1-NA	Teste

Fonte: Autoria própria.

ilustra o funcionamento desses algoritmos na predição dessas proteínas. Nessa representação, os dados X e Y representam informações das proteínas, sendo que os algoritmos aprendem a partir delas e realizam os testes. É possível pensar em X e Y como uma função $f(x) \rightarrow y$. Na fase de teste, o algoritmo utiliza a técnica de validação cruzada (10-*fold cross-validation*), que compara o resultado obtido com o resultado esperado e calcula a precisão do algoritmo na predição das cepas da gripe.

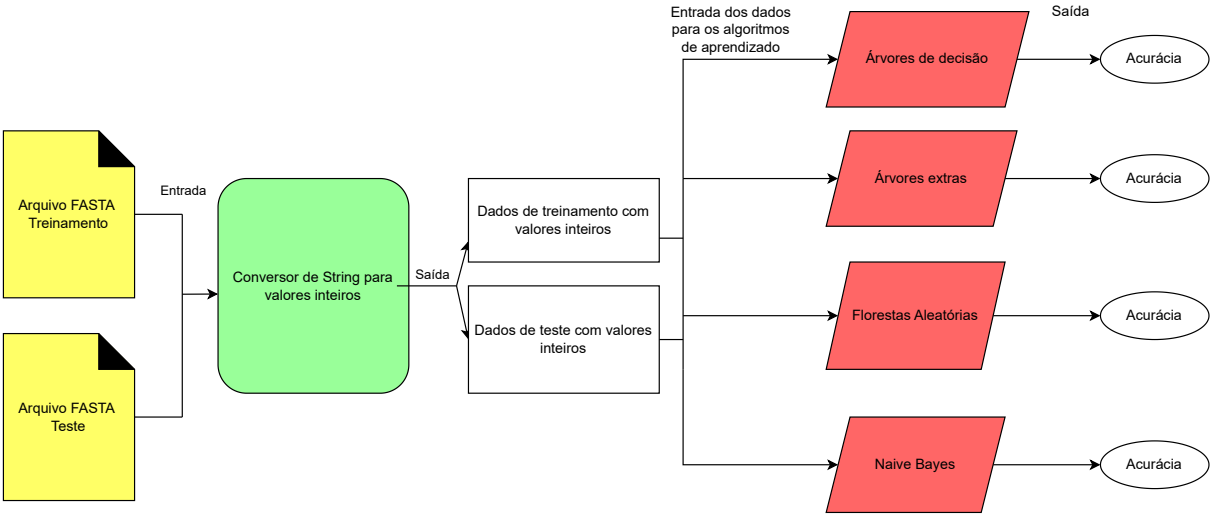
Figura 6 – Representação do funcionamento dos modelos implementados.



Fonte: Adaptado de Koodli (2017).

A partir das figuras 5 e 6 foi possível observar como ocorre partes do processo de predição das glicoproteínas hemaglutina e neuraminidase. Entretanto, para ter uma visão mais ampla de como todo o processo decorre, a figura 7 mostra um esquema completo desde a entrada dos arquivos FASTA de treinamento e de teste, passando pelo conversor até chegar nos modelos de predição, onde ocorre o treinamento e os testes e finalizando com a apresentação da acurácia de cada algoritmo. O código-fonte desenvolvido neste trabalho está disponível em: <https://github.com/lucasnamac/flu-prediction>.

Figura 7 – Representação completa do algoritmo de predição.



Fonte: Autoria própria.

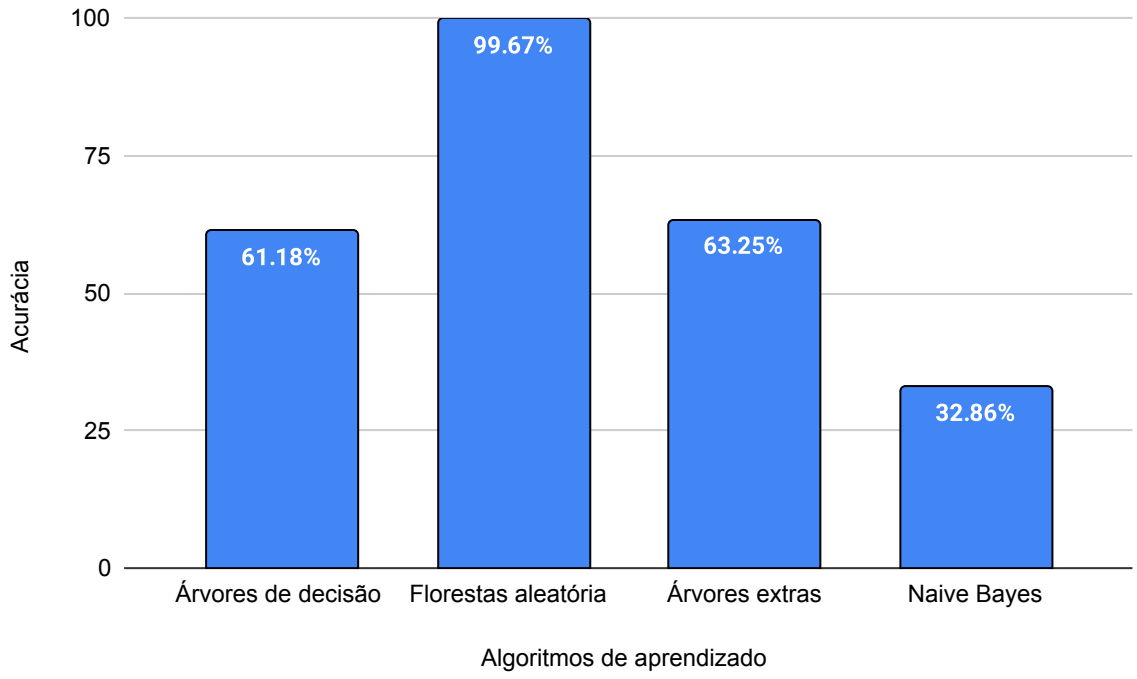
4 Resultados e Discussão

Após a realização de todos os testes, os dados gerados pelos algoritmos foram coletados e armazenados em uma planilha para facilitar a visualização e comparação. Neste capítulo, a apresentação desses desses dados foi dividida em duas partes. A primeira parte consiste na comparação dos algoritmos utilizando os dados de Koodli (2017), permitindo, assim, avaliar o desempenho dos algoritmos com uma base em valores coletada em meados de 2007. A segunda parte utiliza uma base de dados mais recente, correspondente ao período de 2017 a 2022.

4.1 Experimentos com os dados de Koodli (2017)

Na primeira parte, foram utilizadas 100 amostras de hemaglutinina e neuraminidase da cepa H1N1, além de 150 amostras de hemaglutinina da cepa H3N2 e 100 amostras de neuraminidase da cepa H3N2 para treinamento e teste. A Figura 8 apresenta os resultados da predição da proteína hemaglutinina da cepa H1N1 a partir da base de dados de Koodli (2017).

Figura 8 – Resultado da predição da proteína hemaglutinina da cepa H1N1.

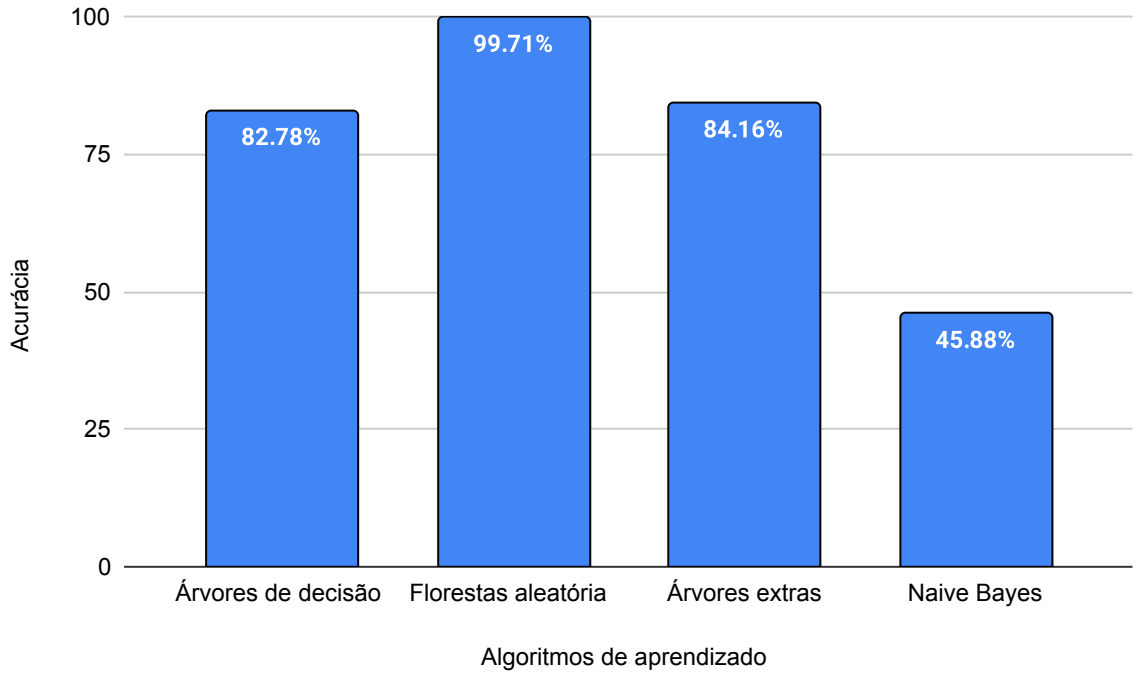


Fonte: Autoria própria

Ao analisar o gráfico, observa-se que o algoritmo de florestas aleatórias obteve um resultado bastante satisfatório, alcançando uma acurácia de 99,67%. Os algoritmos de árvores de decisão e árvores extras apresentaram resultados razoáveis e muito semelhantes, com acurácias de 61,18% e 63,25%, respectivamente. Por fim, o algoritmo *Naive Bayes* teve o desempenho mais baixo, com uma acurácia de apenas 32,86%.

Prosseguindo para o próximo teste com os dados coletados por Koodli (2017), a Figura 9 apresenta o resultado da predição da proteína neuraminidase da cepa H1N1. Analisando o gráfico, observa-se que o algoritmo de florestas aleatórias manteve o melhor desempenho, com uma acurácia de 99,71%, um aumento de 0,04% em relação ao teste anterior. O algoritmo de árvores de decisão apresentou uma melhoria em relação ao teste anterior, alcançando uma acurácia de 82,78%. O algoritmo de árvores extras obteve um desempenho semelhante ao de árvores de decisão, com uma acurácia de 84,16%. Por fim, o algoritmo *Naive Bayes* apresentou uma ligeira melhoria em relação ao teste anterior, mas ainda está longe de fornecer resultados satisfatórios para previsões, com uma acurácia de 45,88%.

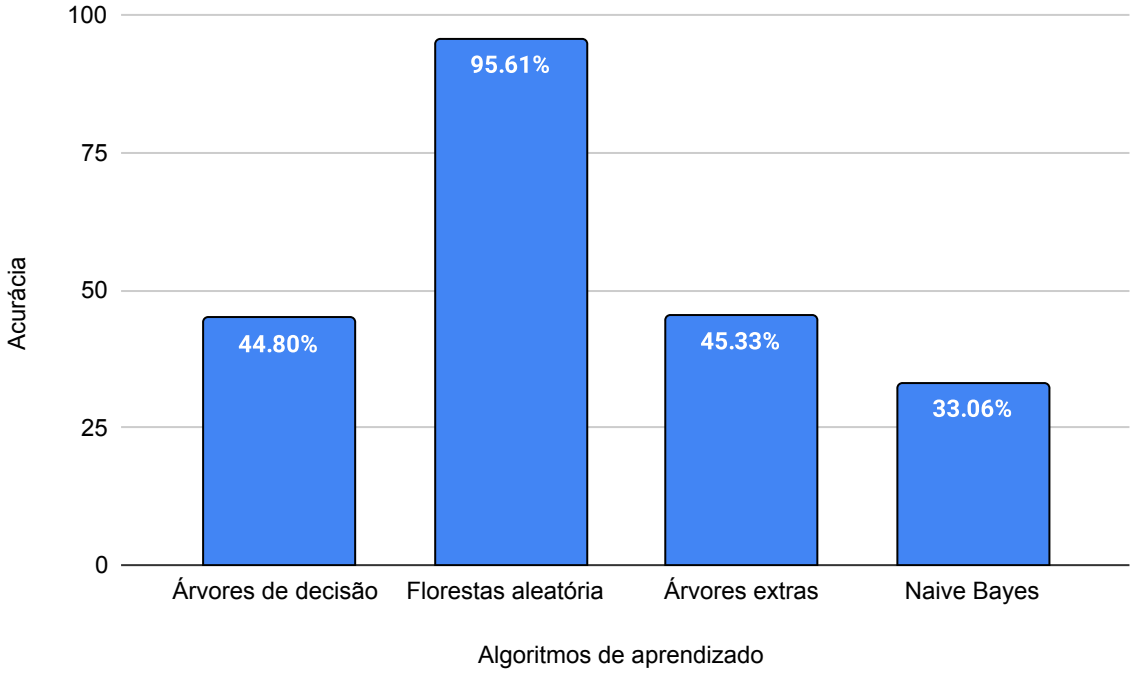
Figura 9 – Resultado da predição da proteína neuraminidase da cepa H1N1.



Fonte: Autoria própria

Os próximos resultados serão referentes às proteínas da cepa H3N2, ainda utilizando os dados coletados em meados de 2007 (KOODLI, 2017). A Figura 10 apresenta os resultados alcançados pelos modelos de predição para a proteína hemaglutinina.

Figura 10 – Resultado da predição da proteína hemaglutinina da cepa H3N2.

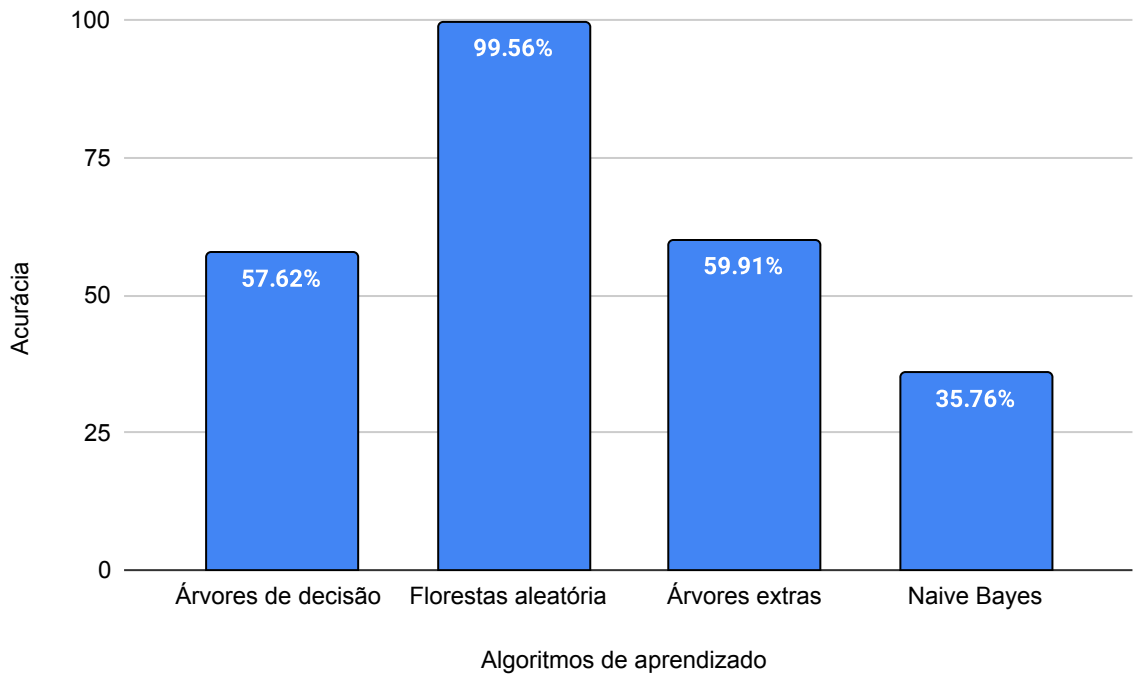


Fonte: Autoria própria

Analisando os dados apresentados na Figura 10, verifica-se que o algoritmo de florestas aleatórias obteve o melhor resultado, com uma acurácia de 95,61%. Por outro lado, os algoritmos de árvores extras, árvores de decisão e *Naive Bayes* não tiveram um bom desempenho neste teste, pois alcançaram uma acurácia inferior a 50%. De maneira mais específica, obtiveram acurácia de 45,33%, 44,80% e 33,06%, respectivamente.

Por fim, ainda usando os dados de Koodli (2017), a Figura 11 apresenta o desempenho dos modelos de predição na previsão da proteína neuraminidase da cepa H3N2. Analisando essa figura, pode-se constatar que o algoritmo de florestas aleatórias obteve um resultado bastante preciso, com uma acurácia de 99,56%. Houve uma pequena melhoria em relação ao teste anterior, um aumento de 3,96%. Os algoritmos de árvores de decisão e árvores extras apresentaram resultados muito próximos entre si, com acurácias de 57,62% e 59,91%, respectivamente. Embora tenham apresentado uma acurácia melhor em comparação ao teste anterior, seus desempenhos ainda são baixos quando comparadas ao modelo florestas aleatórias. Já o algoritmo *Naive Bayes* continuou com o pior desempenho, apresentando um aumento de apenas 2,7% em relação ao teste anterior, chegando a uma acurácia de 35,76%.

Figura 11 – Resultado da predição da proteína neuraminidase da cepa H3N2.

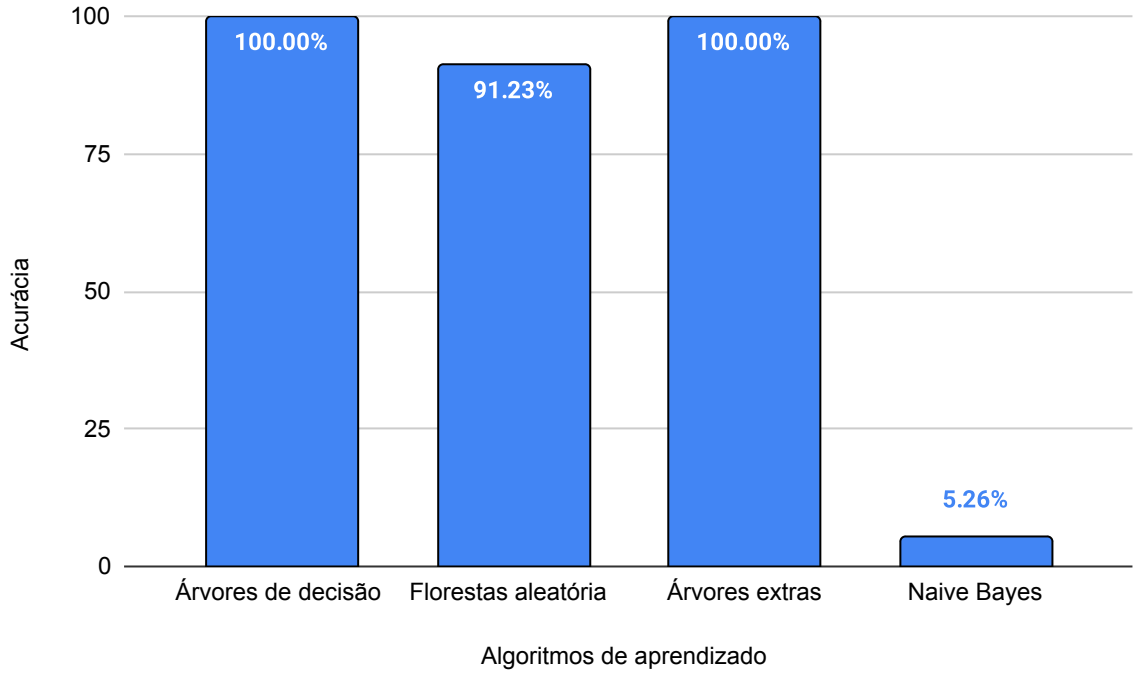


Fonte: Autoria própria

4.2 Experimentos com dados mais recentes

Pela análise dos resultados mostrados na seção anterior, é possível chegar a uma conclusão parcial de que o algoritmo florestas aleatórias apresenta o melhor desempenho para as predições. No entanto, vale ressaltar que os dados utilizados nessa primeira análise são antigos. Para verificar se o algoritmo de florestas aleatórias continua sendo o melhor para as predições do vírus influenza, foi realizado um segundo experimento utilizando cepas mais recentes. A Figura 12 apresenta os resultados obtidos pelos algoritmos a partir dessas cepas mais recentes, obtidas no período de 2017 a 2022.

Figura 12 – Resultado da predição da proteína hemaglutinina da cepa H1N1 usando dados mais recentes.

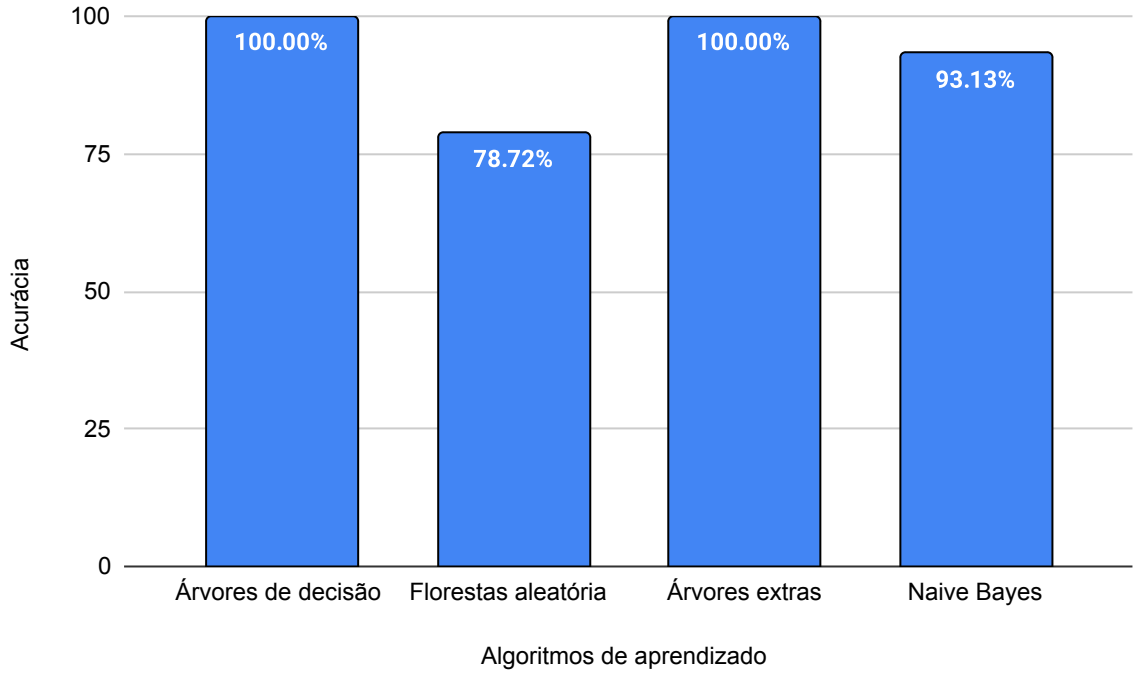


Fonte: Autoria própria

Interpretando os dados apresentados na Figura 12, observa-se que, diferentemente dos resultados anteriores, neste experimento os algoritmos que apresentaram o melhor desempenho foram árvores aleatórias e árvores extras, ambos com 100% de acurácia na predição da proteína hemaglutinina da cepa H1N1. O algoritmo de florestas aleatórias, que havia se destacado como o melhor anteriormente, também obteve um bom desempenho, com uma acurácia de 91,23%. Por outro lado, o algoritmo *Naive Bayes* teve o pior desempenho até o momento, com apenas 5,26% de acurácia.

Continuando a apresentação dos resultados, a Figura 13 mostra os resultados na predição da proteína neuraminidase da cepa H1N1. É possível constatar que todos os modelos de predição tiveram um bom desempenho na previsão dessa proteína. Os algoritmos de árvores de decisão e árvores extras, assim como no experimento anterior, apresentaram o melhor desempenho, alcançando 100% de acurácia. O algoritmo *Naive Bayes* também teve um bom desempenho, com uma taxa de predição de 93,13%. Finalmente, o algoritmo de florestas aleatórias obteve uma acurácia de 78,72%, o que também é considerado um bom valor.

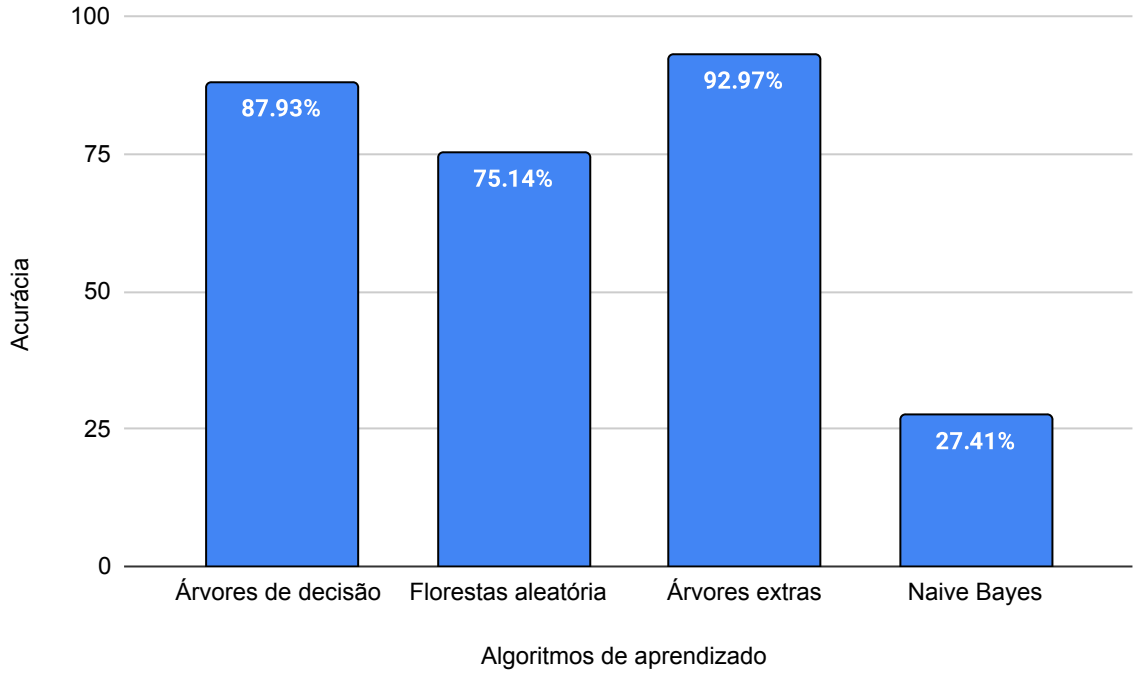
Figura 13 – Resultado da predição da proteína neuraminidase da cepa H1N1 usando dados mais recentes.



Fonte: Autoria própria

Para finalizar a apresentação dos testes, será mostrado agora os resultados obtidos para as proteínas hemaglutinina e neuraminidase referentes a cepa H3N2. A figura 14 exibe os resultados na predição da proteína hemaglutinina.

Figura 14 – Resultado da predição da proteína hemaglutinina da cepa H3N2 usando dados mais recentes.

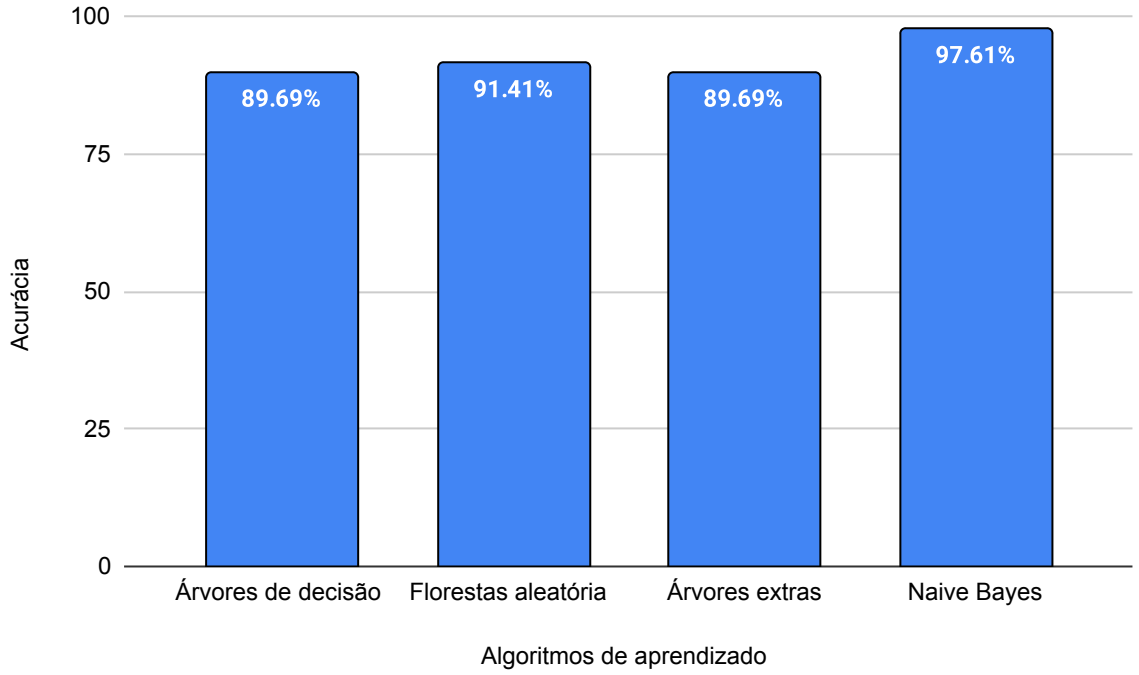


Fonte: Autoria própria

Analisando a figura 14 podemos compreender que o algoritmo que teve o melhor desempenho nesse experimento foi árvore extra, ficando com uma acurácia de 92,97%. Logo em seguida, outro algoritmo que teve um bom desempenho foi árvores de decisão com uma acurácia de 87,93%. O terceiro algoritmo teve que teve um desempenho muito bom foi floresta aleatória com uma acurácia de 75,14%. Já o algoritmo *Naive Bayes* teve novamente um baixo desempenho ficando apenas com 27,41%.

Por fim, a Figura 15 apresenta os resultados do último experimento realizado, que trata da predição da proteína neuraminidase da cepa H3N2. No último caso, todos os modelos de predição obtiveram resultados bastante satisfatórios. O algoritmo *Naive Bayes* se destacou, alcançando uma acurácia de 97,61%. Em segundo lugar, o algoritmo floresta aleatória apresentou um bom desempenho, com uma acurácia de 91,41%. Os algoritmos de árvores de decisão e árvores extras tiveram exatamente a mesma precisão na predição, atingindo 89,69%, o que também pode ser considerado um bom resultado.

Figura 15 – Resultado da predição da proteína neuraminidase da cepa H3N2 usando dados mais recentes.



Fonte: Autoria própria

4.3 Comparação com resultados da literatura

Nas seções anteriores, foram apresentados os resultados obtidos durante a realização dos testes, e, a partir de uma análise desses dados, é possível inferir que, de forma geral, o algoritmo florestas aleatórias se mostrou o melhor para a predição. Em todos os testes realizados, sua acurácia se manteve sempre alta, não ficando abaixo de 75% em nenhum cenário. Outros modelos de predição apresentaram bons desempenhos em determinados casos, mas tiveram um desempenho consideravelmente ruim em outros, como é o caso do algoritmo de árvores extras, que obteve apenas 45,33% de acurácia na predição da proteína hemaglutinina da cepa H3N2.

Por último, é necessário fazer uma comparação entre os resultados obtidos usando o algoritmo florestas aleatórias e os resultados encontrados na literatura. Essa comparação é importante para analisar se um algoritmo de aprendizado de máquina pode ser um recurso adicional para a detecção de novas cepas do vírus influenza. Para essa análise, coletou-se todos os resultados gerados pelo algoritmo floresta aleatória e computou-se a média aritmética e o desvio padrão da acurácia. A média aritmética é utilizada diretamente na comparação com outros modelos de predição presentes na literatura, enquanto o desvio padrão é calculado para conferir maior confiabilidade à média aritmética. A Tabela 4

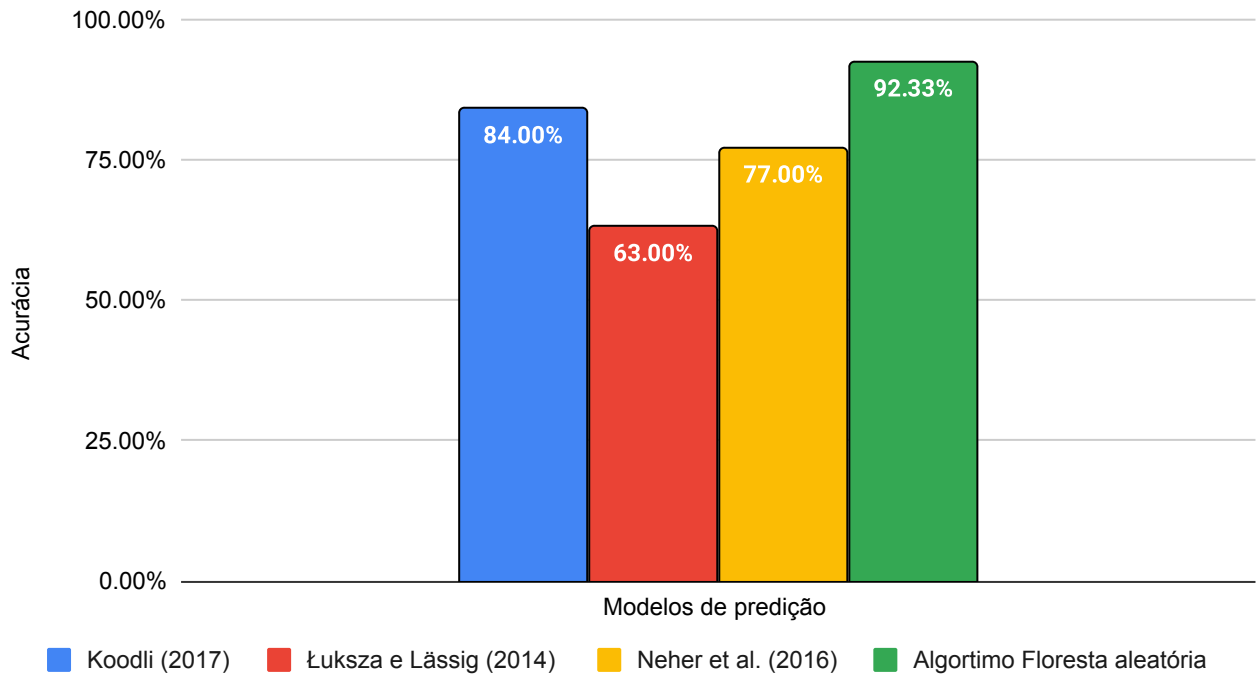
apresenta o cálculo da média e do desvio padrão para cada cepa.

Tabela 4 – Cálculo da média e desvio padrão dos resultados obtidos nos experimentos.

Cepa	Acurácia Média	Desvio Padrão
H1N1	92,33	0,099
H3N2	90,43	0,10

A partir dos dados apresentados na Tabela 4, é possível comparar os resultados encontrados para as cepas H1N1 e H3N2 com os resultados encontrados em outros estudos da literatura. A Figura 16 mostra uma comparação dos estudos realizados para a predição da cepa H1N1 e os resultados deste trabalho.

Figura 16 – Comparação dos resultados da predição da cepa H1N1 deste trabalho com os presentes na literatura.

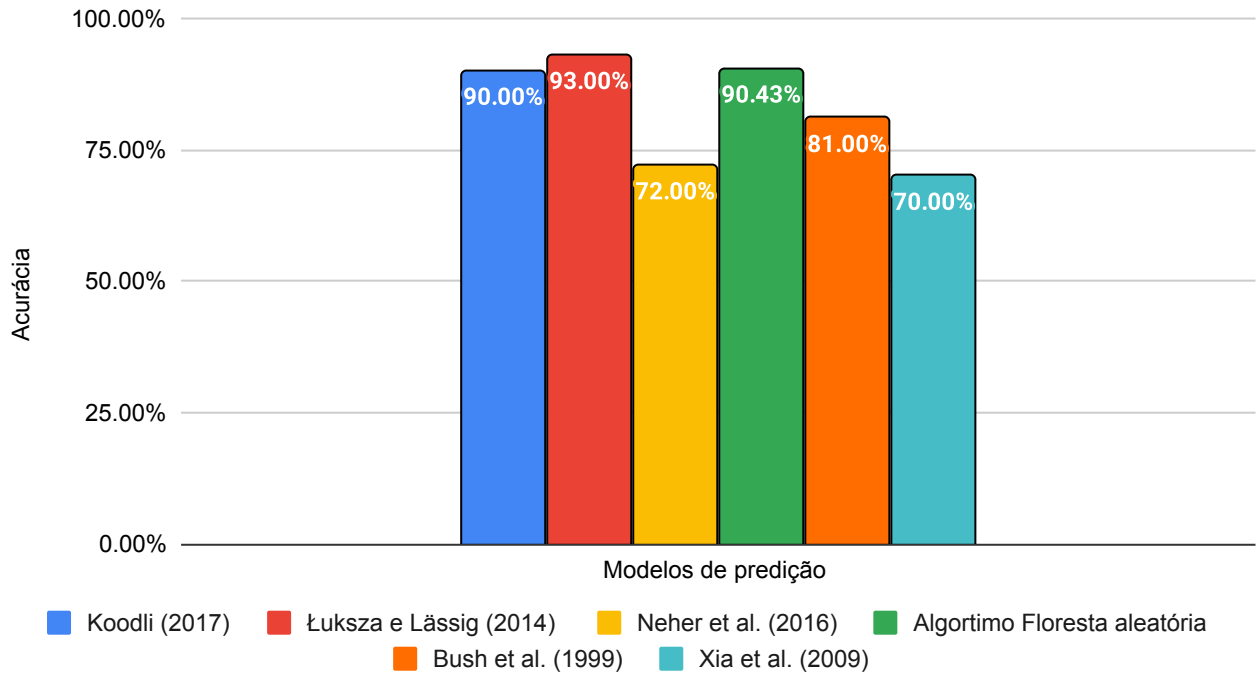


Fonte: Autoria própria

Ao examinarmos a Figura 16, é possível constatar que o algoritmo florestas aleatórias apresentou um bom desempenho quando comparado a outros métodos de predição presentes na literatura. Portanto, nesse contexto, pode-se concluir que para a predição da cepa H1N1 do vírus influenza, o algoritmo florestas aleatórias é o melhor método de predição.

Por fim, na Figura 17, são comparados os resultados obtidos pelo algoritmo florestas aleatórias com os estudos encontrados na literatura. A partir dessa comparação, observamos que o algoritmo florestas aleatórias não foi o melhor na predição da cepa H3N2. No entanto, obteve a segunda melhor predição, superando diversos métodos. Portanto, é possível concluir que o algoritmo florestas aleatórias é um excelente método para a predição da cepa H3N2, e, juntamente com os métodos já existentes, pode contribuir ainda mais para o desenvolvimento de novas vacinas, proporcionando um impacto significativamente benéfico para a sociedade.

Figura 17 – Comparação dos resultados da predição da cepa H3N2 deste trabalho com os presentes na literatura.



Fonte: Autoria própria

É importante ressaltar que os experimentos descritos pela literatura não foram reproduzidos e tampouco a confrontação foi feita dentro de um mesmo ambiente de comparação. A comparação entre o algoritmo florestas aleatória com os modelos presentes na literatura não buscou garantir igualdade de condições para todos eles, apenas foram coletados os resultados mencionado pelos trabalhos já existentes e comparados com o algoritmo de florestas aleatórias proposto neste trabalho.

5 Conclusões

Todos os anos, são realizados estudos para tentar desenvolver novas vacinas ou medicamentos que sejam eficazes contra o vírus influenza. No entanto, essa classe de vírus está sempre sofrendo mutações, o que dificulta o desenvolvimento de um tratamento eficiente. Nesse contexto, Este trabalho de conclusão propôs uma comparação de algoritmos de aprendizado de máquina na predição de novas cepas dos vírus H1N1 e H3N2 com o objetivo de eleger o algoritmo mais adequado para fazer esse tipo de previsão. Os resultados mostraram que, dentre os modelos avaliados, o melhor para esse tipo de predição é o Floresta Aleatória. Quando esse mesmo algoritmo foi comparado a modelos descritos na literatura, conseguiu alcançar uma acurácia superior a quatro outros modelos já existentes. Isso indica que a inteligência artificial pode possibilitar um avanço no desenvolvimento de vacinas contra agentes infecciosos de forma mais acelerada, promovendo a proteção da população.

Nesse sentido, é interessante que sejam empregados outros modelos de aprendizado, a fim de garantir o desenvolvimento de vacinas eficientes e eficazes. Trabalhos futuros podem usar modelos como *Gradient Boosting Machines* (GBM) e regressão linear para fazer a predição de novos vírus. De acordo com Natekin e Knoll (2013), a ideia principal do modelo GBM é construir modelos de aprendizados base com um valor gradiente associado e os resultados da predição desses modelos são correlacionados com o valor gradiente da função de perda. Já o modelo de regressão linear, segundo Maulud e Abdulazeez (2020), consiste em avaliar e quantificar a relação entre as variáveis. Pham et al. (2020) fizeram uma comparação do desempenho do algoritmo florestas aleatórias e redes neurais artificiais, mostrando que ambos possuem um grande potencial na predição. Dessa forma, uma outra linha de pesquisa que pode ser seguida é usar diferentes modelos redes neurais para tentar prever novas cepas.

Finalmente, é importante destacar que, segundo a Organização Mundial da Saúde (OMS), novas variantes de vírus, como a influenza e o coronavírus, podem surgir com um potencial mais letal do que as já conhecidas (ONU, 2023). Além disso, também existe a ameaça de surgimento de agentes patógenos com potencial de letalidade ainda maior, destacando a importância do monitoramento e de estudos para o desenvolvimento de vacinas.

Assim, é fundamental que o desenvolvimento e aplicação de algoritmos de aprendizado de máquina para prever não somente cepas de H1N1 e H3N2, mas diversos outros vírus, principalmente aqueles responsáveis por surtos epidêmicos ao longo da história (como ebola, coronavírus e zikavírus). Monitorar a evolução desses agentes patológicos é

de suma importância do ponto de vista econômico, social e sanitário (ONU, 2023).

Referências

AHMAD, M. W.; REYNOLDS, J.; REZGUI, Y. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. **Journal of Cleaner Production**, Elsevier BV, v. 203, p. 810–821, dez. 2018. DOI: <https://doi.org/10.1016/j.jclepro.2018.08.207>. Citado na página 21.

ARA-SOUZA, A. L. **Redes Bayesianas: Uma introdução aplicada a credit scoring**. 99 p. Monografia (Graduação) — Universidade Federal de São Carlos, São Carlos, SP, 2010. Citado na página 17.

BIAU, G.; SCORNET, E. A random forest guided tour. **TEST**, Springer Science and Business Media LLC, v. 25, n. 2, p. 197–227, abr. 2016. DOI: <https://doi.org/10.1007/s11749-016-0481-7>. Citado na página 19.

BRASIL. Ministério da Saúde. **Doenças infecciosas e parasitárias: Guia de bolso**. Brasília, DF, 2010. 442 p. Citado na página 23.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001. DOI: <https://doi.org/10.1023/A:1010933404324>. Citado 2 vezes nas páginas 19 e 20.

BUSH, R. M.; BENDER, C. A.; SUBBARAO, K.; COX, N. J.; FITCH, W. M. Predicting the evolution of human influenza A. **Science**, American Association for the Advancement of Science (AAAS), v. 286, n. 5446, p. 1921–1925, dez. 1999. DOI: <https://doi.org/10.1126/science.286.5446.1921>. Citado na página 26.

CARTER, J.; SAUNDERS, V. **Virology**: Principles and applications. 2. ed. [S.l.]: Wiley & Sons, Limited, John, 2013. 394 p. ISBN 978-1119991427. Citado na página 23.

COMONE, P. **Diversidade genética da hemaglutinina (HA) de vírus influenza A, entre 1995 e 2006**. 134 p. Tese (Doutorado) — Universidade de São Paulo, São Paulo, set. 2011. DOI: <https://doi.org/10.11606/d.87.2011.tde-27092011-104745>. Citado 2 vezes nas páginas 12 e 13.

DAUMÉ III, H. **A course in machine learning**. Maryland: Self-published, 2017. 227 p. Citado 3 vezes nas páginas 16, 18 e 20.

FERNANDES, A. M. d. R. **Inteligência Artificial**: Noções gerais. Rio de Janeiro, RJ: Visual Books, 2004. 146 p. ISBN 978-8575021149. Citado na página 15.

FRANCO, C. R. **Inteligência Artificial**. Indaial, SC: Editora UNIASSELVI, 2014. 180 p. ISBN 855-1500668. Citado na página 15.

FRANK, E.; TRIGG, L.; HOLMES, G.; WITTEN, I. H. Technical note: Naïve bayes for regression. **Machine Learning**, Springer Science and Business Media LLC, v. 41, n. 1, p. 5–25, 2000. DOI: <https://doi.org/10.1023/a:1007670802811>. Citado na página 30.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. **Machine Learning**, Springer Science and Business Media LLC, v. 63, n. 1, p. 3–42, mar. 2006. DOI: <https://doi.org/10.1007/s10994-006-6226-1>. Citado 2 vezes nas páginas 21 e 22.

GIBAS, C.; JAMBECK, P. **Developing Bioinformatics Computer Skills**: An introduction to software tools for biological applications. EUA: O'Reilly Media, 2001. 427 p. ISBN 978-1565926646. Citado na página 29.

GOMES, D. V. **Variantes do Vírus Influenza: Prevenção e Tratamento**. 27 p. Dissertação (Mestrado) — Universidade de Coimbra, Coimbra, set. 2014. Disponível em: <https://estudogeral.uc.pt/handle/10316/92639>. Acesso em: 27 jun. 2023. Citado 2 vezes nas páginas 12 e 16.

GRANATO, C. F. H.; BELLEI, N. C. J. As novas facetas e a ameaça da gripe aviária no mundo globalizado. **Jornal Brasileiro de Patologia e Medicina Laboratorial**, FapUNIFESP (SciELO), v. 43, n. 4, p. 245–249, ago. 2007. DOI: <https://doi.org/10.1590/s1676-24442007000400005>. Citado na página 12.

JIANG, L.; ZHANG, H.; CAI, Z. A novel bayes model: Hidden naive bayes. **IEEE Transactions on Knowledge and Data Engineering**, Institute of Electrical and Electronics Engineers (IEEE), v. 21, n. 10, p. 1361–1371, out. 2009. DOI: <https://doi.org/10.1109/tkde.2008.234>. Citado na página 17.

KLINGEN, T. R.; REIMERING, S.; LOERS, J.; MOOREN, K.; KLAWONN, F.; KREY, T.; GABRIEL, G.; MCHARDY, A. C. Sweep dynamics (SD) plots: Computational identification of selective sweeps to monitor the adaptation of influenza a viruses. **Scientific Reports**, Springer Science and Business Media LLC, v. 8, n. 1, p. 1–13, jan. 2018. DOI: <https://doi.org/10.1038/s41598-017-18791-z>. Citado na página 27.

KOODLI, R. **Applying Machine Learning to Study Influenza Virus Behavior**. 2017. Disponível em: <https://github.com/RK900/Flu-Prediction>. Acesso em: 30 jun. 2022. Citado 7 vezes nas páginas 13, 25, 29, 32, 34, 35 e 36.

LEES, W. D.; MOSS, D. S.; SHEPHERD, A. J. A computational analysis of the antigenic properties of haemagglutinin in influenza a h3n2. **Bioinformatics**, Oxford University Press (OUP), v. 26, n. 11, p. 1403–1408, abr. 2010. DOI: <https://doi.org/10.1093/bioinformatics/btq160>. Citado na página 26.

ŁUKSZA, M.; LäSSIG, M. A predictive fitness model for influenza. **Nature**, Springer Science and Business Media LLC, v. 507, n. 7490, p. 57–61, fev. 2014. DOI: <https://doi.org/10.1038/nature13087>. Citado na página 25.

MARTINS, E. G. M. Coeficiente de determinação. **Revista Ciência Elementar**, v. 6, n. 1, p. 24, 2018. DOI: <http://doi.org/10.24927/rce2018.024>. Citado na página 30.

MAULUD, D.; ABDULAZEEZ, A. M. A review on linear regression comprehensive in machine learning. **Journal of Applied Science and Technology Trends**, v. 1, n. 4, p. 140–147, 2020. DOI: <https://doi.org/10.38094/jastt1457>. Citado na página 44.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. In: REZENDE, S. O. (Ed.). **Sistemas Inteligentes: Fundamentos e Aplicações**. 1. ed. Barueri, SP: Manole Ltda, 2003. p. 89–114. ISBN 85-204168. Citado na página 16.

MORAIS, R. L. de. **Uso de árvores aleatórias para classificação sensorial de arroz cozido**. 70 p. Monografia (Graduação) — Universidade de Brasília, Brasília, DF, 2017. Citado na página 20.

NATEKIN, A.; KNOLL, A. Gradient boosting machines, a tutorial. **Frontiers in neurorobotics**, Frontiers Media SA, v. 7, p. 21, 2013. DOI: <https://doi.org/10.3389/fnbot.2013.00021>. Citado na página 44.

NEHER, R. A.; BEDFORD, T.; DANIELS, R. S.; RUSSELL, C. A.; SHRAIMAN, B. I. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. **Proceedings of the National Academy of Sciences**, Proceedings of the National Academy of Sciences, v. 113, n. 12, p. E1701–E1709, mar. 2016. DOI: <https://doi.org/10.1073/pnas.1525578113>. Citado na página 25.

NILSSON, N. J. **Introdution to machine learning**. Stanford, CA: Stanford University, 1996. 197 p. Citado 2 vezes nas páginas 16 e 18.

NOGUEIRA, T.; PONCE, R. O vírus da gripe. **Revista de Ciência Elementar**, ICETA, v. 9, n. 2, p. 1–5, jun. 2021. DOI: <https://doi.org/10.24927/rce2021.038>. Citado na página 12.

NOSSEDOTTI. **Replication cycle of a hypothetical virus**. 2011. Disponível em: <https://commons.wikimedia.org/wiki/File:Virus_replication.svg>. Acesso em: 31 maio 2023. Citado na página 24.

_____. **Scheme of unenveloped and enveloped virions**. 2011. Disponível em: <https://commons.wikimedia.org/wiki/File:Virus_scheme.svg>. Acesso em: 31 maio 2023. Citado na página 24.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS. **World must be ready to respond to next pandemic: WHO chief**. 2023. Disponível em: <https://news.un.org/en/story/2023/05/1136912>. Acesso em: 31 maio 2023. Citado 2 vezes nas páginas 44 e 45.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Citado na página 29.

PENNINGS, P. S.; HERMISSON, J. Soft sweeps II – molecular population genetics of adaptation from recurrent mutation or migration. **Molecular Biology and Evolution**, Oxford University Press (OUP), v. 23, n. 5, p. 1076–1084, mar. 2006. DOI: <https://doi.org/10.1093/molbev/msj117>. Citado na página 27.

PHAM, T. A.; LY, H.-B.; TRAN, V. Q.; GIAP, L. V.; VU, H.-L. T.; DUONG, H.-A. T. Prediction of pile axial bearing capacity using artificial neural network and random forest. **Applied Sciences**, MDPI, v. 10, n. 5, p. 1871, 2020. DOI: <https://doi.org/10.3390/app10051871>. Citado na página 44.

RUSSEL, S.; NORVIG, P. **Inteligência Artificial**. Rio de Janeiro, RJ: Elsevier, 2013. 1323 p. ISBN 978-8535237016. Citado 2 vezes nas páginas 15 e 16.

SAXENA, S. K.; KOTIKALAPUDI, R.; TIWARI, S.; MUVVA, C. Influenza A(H1N1)pdm09 virus: therapeutics and challenges. **Future Virology**, Future Medicine Ltd, v. 7, n. 10, p. 947–950, out. 2012. DOI: <https://doi.org/10.2217/fvl.12.90>. Citado na página 12.

SCHWARCZ, L. M.; STARLING, H. M. **A bailarina da morte: a gripe espanhola no Brasil**. São Paulo, SP: Companhia das Letras, 2020. 368 p. ISBN 978-8535933918. Citado na página 12.

SILVA FILHO, E. B. da; SILVA, A. L. da; SANTOS, A. de Oliveira dos; DALL'ACQUA, D. S. V.; SOUZA, L. F. B. Infecções respiratórias de importância clínica: Uma revisão sistemática. **Revista Faculdades Integradas Aparício Carvalho**, v. 4, n. 1, p. 7–16, dez. 2017. Disponível em: <https://www.arca.fiocruz.br/handle/icict/33445>. Acesso em: 27 jun. 2023. Citado na página 12.

SUZUKI, Y. Predictability of antigenic evolution for h3n2 human influenza a virus. **Genes & Genetic Systems**, Genetics Society of Japan, v. 88, n. 4, p. 225–232, 2013. DOI: <https://doi.org/10.1266/ggs.88.225>. Citado 2 vezes nas páginas 26 e 27.

XIA, Z.; JIN, G.; ZHU, J.; ZHOU, R. Using a mutual information-based site transition network to map the genetic evolution of influenza a/h3n2 virus. **Bioinformatics**, Oxford University Press (OUP), v. 25, n. 18, p. 2309–2317, ago. 2009. DOI: <https://doi.org/10.1093/bioinformatics/btp423>. Citado na página 28.

YONG, E. Scientists create hybrid flu that can go airborne. **Nature**, Springer Science and Business Media LLC, maio 2013. DOI: <https://doi.org/10.1038/nature.2013.12925>. Citado na página 12.