UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
CENTRO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO

Substituições de Aminoácidos como Representação de Mutações em Transformers para Análise e Classificação de Variantes do SARS-CoV-2

Alessandro Soares da Silva

Orientador: Prof. Dr. Marcelo Augusto Costa Fernandes

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Engenharia Elétrica e de Computação da UFRN. (área de concentração: Engenharia de Computação) como parte dos requisitos para a obtenção do título de Mestre em Ciências.

Numero de ordem PPgEEC: M698
Natal, RN, Dezembro 2025

Universidade Federal do Rio Grande do Norte - UFRN

Sistema de Bibliotecas - SISBI

Catalogação de Publicação na Fonte. UFRN - Biblioteca Central Zila Mamede

Silva, Alessandro Soares da.

Substituições de aminoácidos como representação de mutações em transformers para análise e classificação de variantes do SARS-CoV-2 / Alessandro Soares da Silva - 2025.

84 f.: il.

Dissertação (mestrado) - Universidade Federal do Rio Grande do Norte, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica e de Computação, Natal, RN, 2025.

Orientação: Prof. Dr. Marcelo Augusto Costa Fernandes.

1. Inteligência Artificial - Dissertação. 2. SARS-CoV-2 - Dissertação. 3. Substituições de Aminoácidos - Dissertação. 4. Embeddings - Dissertação. 5. Transformers - Dissertação. 6. Clusterização Fuzzy - Dissertação. I. Fernandes, Marcelo Augusto Costa. II. Título.

RN/UF/BCZM CDU 004.8:578.834

Alessandro Soares da Silva

Dissertação de Mestrado

Prof. Dr. Marcelo Augusto Costa Fernandes (orientador) ..... DCA/UFRN

Prof. Dr. Gabriel Bezerra Motta Câmara ..... InovAI Lab/Lance/nPiti/UFRN

Prof. Dr. Patrick Cesar Alves Terrematte ..... IMD/UFRN

Profa. Dra. Raquel de Melo Barbosa ..... UGR

Aos meus filhos, Latifah e Kasem
Neto, por todo carinho e amor dado
a mim.

Agradecimentos

Aos meus filhos Latifah e Kasem Neto, que foram e são a razão da minha força de vontade e inspiração durante toda essa caminhada.

Para Rechfy Kasem, meu amor, que me ajudou, aconselhou e aguentou durante todas as vitórias e tristeza que essa etapa da minha vida trouxe. Você fez tornar essa caminhada mais fácil.

Ao meu pai que sempre deu seu apoio incondicional, e a minha mãe, que com toda certeza está torcendo por mim lá do céu.

Ao meu orientador, pela paciência e por ter me mostrado os caminhos corretos para o sucesso acadêmico.

Aos colegas de pós-graduação, em especial do InovAI, pelas conversas, sugestões e trocas de ideias, que agregaram muito ao meu aprendizado.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), pelo apoio financeiro.

Resumo

A pandemia de COVID-19, causada pelo SARS-CoV-2, evidenciou necessidade de métodos computacionais escaláveis e automatizáveis para vigilância genômica frente ao surgimento contínuo de variantes de preocupação. Neste contexto, este trabalho propõe uma abordagem inovadora baseada em representações vetoriais de substituições de aminoácidos, explorando modelos *Transformer* para a análise e classificação de variantes virais do SARS-CoV-2, com foco nas variantes Alpha, Beta, Gamma, Delta e Omicron. As sequências de substituições de aminoácidos foram processadas em *embeddings* de alta dimensionalidade, os quais fundamentaram dois experimentos complementares. No primeiro, aplicaram-se técnicas não supervisionadas, como Fuzzy C-Means e projeção t-SNE, capazes de revelar agrupamentos coerentes com as variantes conhecidas, além de identificar zonas de transição e amostras ambíguas. No segundo, modelos de aprendizado supervisionado, incluindo SVM, Random Forest, k-NN e XGBoost, foram avaliados na tarefa de classificação de varintes, com destaque para o XGBoost, que alcançou acurácia superior a 99,83% em *dataset* externo de teste. Os resultados demonstram que representações derivadas exclusivamente de substituições de aminoácidos são capazes de discriminar variantes com elevado desempenho e capturar assinaturas mutacionais de relevância biológica, sem necessidade de alinhamento genômico. Desta forma, a abordagem proposta apresenta-se como uma alternativa escalável, flexível, adaptável e puramente computacional para a vigilância automatizada de patógenos, com potencial aplicação em cenários de saúde pública.

Palavras-chave: SARS-CoV-2; Inteligência Artificial; Substituições de Aminoácidos; *Embeddings*; *Transformers*; Clusterização Fuzzy; Classificação Supervisionada; Vigilância Genômica.

Abstract

The COVID-19 pandemic, caused by SARS-CoV-2, highlighted the need for scalable and automatable computational methods for genomic surveillance in the face of the continuous emergence of variants of concern. In this context, this work proposes an innovative approach based on vector representations of amino acid substitutions, leveraging Transformer models for the analysis and classification of SARS-CoV-2 viral variants, with a focus on the Alpha, Beta, Gamma, Delta, and Omicron variants. Sequences of amino acid substitutions were processed into high-dimensional embeddings, which supported two complementary experiments. In the first, unsupervised techniques, such as Fuzzy C-Means clustering and t-SNE projection, were applied to reveal groupings consistent with known variants, as well as transitional regions and ambiguous samples. In the second, supervised learning models, including SVM, Random Forest, k-NN, and XGBoost, were evaluated for variant classification, with XGBoost achieving accuracy above 99% on an external dataset. The results demonstrate that representations derived exclusively from amino acid substitutions are capable of discriminating viral variants with high performance and capturing mutational signatures of biological relevance, without requiring genomic alignment. Thus, the proposed approach represents a scalable, flexible, adaptable, and purely computational alternative for automated pathogen genomic surveillance, with potential applications in public health scenarios.

Keywords: SARS-CoV-2; Artificial Intelligence; Amino Acid Substitutions; Embeddings; Transformers; Fuzzy Clustering; Supervised Classification; Genomic Surveillance.

Sumário

Sumário i

Lista de Figuras iii

Lista de Tabelas v

Lista de Acrônimos e Abreviações vi

1 Introdução 1
- 1.1 A Pandemia de COVID-19: Panorama Global e Impactos . . . 1
- 1.2 Estrutura do SARS-CoV-2 e Mecanismo de Infecção . . . 2
- 1.3 Estado da Arte . . . 5
- 1.4 Objetivos da Pesquisa . . . 7
  - 1.4.1 Objetivo Geral . . . 8
  - 1.4.2 Objetivos Específicos . . . 8
- 1.5 Organização do Trabalho . . . 8
- 1.6 Trabalho Aceitos em Congressos . . . 9

2 Fundamentação Teórica 11
- 2.1 Fundamentos de PLN e Representações Semânticas . . . 11
  - 2.1.1 Tokenização: Transformando Sequências em Unidades de Entrada 11
  - 2.1.2 Embeddings: Representações Vetoriais Densas e Contextuais . . . 12
  - 2.1.3 Arquitetura Transformer: Fundamentos e Mecanismos . . . 14
  - 2.1.4 Modelos Pré-Treinados e Transferência para Domínio Biológico . 16
  - 2.1.5 Modelo All-MiniLM-L6-v2 . . . 17
- 2.2 Redução de Dimensionalidade em Espaços de Embeddings . . . 17
  - 2.2.1 t-SNE: Teoria e Funcionamento . . . 18
- 2.3 Técnicas de Clusterização para Análise de Embeddings . . . 19
  - 2.3.1 K-means Clustering . . . 20
  - 2.3.2 Fuzzy C-means Clustering . . . 21
- 2.4 Aprendizado Supervisionado . . . 22
  - 2.4.1 Modelos Supervisionados para Classificação de Variantes . . . 22
  - 2.4.2 Modelos Baseados em Ensemble . . . 23
  - 2.4.3 Considerações sobre Escolha de Algoritmo . . . 24
- 2.5 Métricas de Avaliação . . . 24
  - 2.5.1 Acurácia . . . 25

2.5.2 Logarithmic Loss (LogLoss) . . . 25
2.5.3 F1-Macro . . . 26
2.5.4 Métricas Complementares de Avaliação . . . 27

3 Proposta, Métodos e Experimentos 28
3.1 Proposta . . . 28
3.2 Metodologia . . . 29
3.2.1 Representação AASubstitutions . . . 30
3.2.2 Geração de Embeddings . . . 31
3.3 Experimentos . . . 34
3.3.1 Agrupamento com o K-means . . . 35
3.3.2 Agrupamento com Fuzzy C-means . . . 35
3.3.3 Redução de Dimensionalidade com t-SNE . . . 38
3.3.4 Analise Supervisionada: Classificação . . . 40

4 Resultados e Discussões 42
4.1 Resultados com Base nos Agrupamentos K-Means . . . 42
4.2 Resultados com Base nos Agrupamentos Fuzzy C-Means . . . 44
4.2.1 Caracterização e Validação dos Agrupamentos Fuzzy . . . 45
4.2.2 Identificação de Amostras Ambíguas e Perfis Transicionais . . . 46
4.2.3 Determinação de Assinaturas Mutacionais por Exclusividade Fuzzy 47
4.3 Resultados com Base no Desempenho de Modelos Supervisionados para Classificação de Variantes . . . 50
4.3.1 As Limitações do SVM na Classificação de Variantes . . . 50
4.3.2 Modelos Baseados em Árvores de Decisão: XGBoost . . . 51
4.3.3 Desempenho do Modelo Random Forest . . . 53
4.3.4 O Desempenho do k-NN e suas Implicações . . . 55
4.4 Desempenho Comparativo dos classificadores e Validação em Dados Externos . . . 56

5 Conclusão 59
5.1 Trabalhos Futuros . . . 60

Referências Bibliográficas 61

Lista de Figuras

1.1 Representação estrutural do SARS-CoV-2, destacando seus principais componentes virais e proteínas estruturais. Adaptada de (Ali, 2024) 3
1.2 Estrutura do genoma do SARS-CoV-2, evidenciando a organização das regiões codificadoras e não codificadoras, bem como as principais proteínas virais associadas. Adaptada de (Fang et al., 2023) 4
1.3 Diagrama esquemático da interação entre a proteína Spike do SARS-CoV-2 e o receptor ACE2 da célula hospedeira, ilustrando o mecanismo molecular de entrada viral. Adaptada de (Villoutreix et al., 2021) 4

2.1 Representação vetorial da entrada no BERT, composta pela soma dos *embeddings* de *token*, segmento e posição, que conjuntamente codificam informações semânticas, estruturais e posicionais da sequência. Adaptado de (Devlin et al., 2018) 13
2.2 A arquitetura geral do modelo *Transformer*, ilustrada em (a), com camadas empilhadas de autoatenção e redes totalmente conectadas. Em (b), apresenta-se o funcionamento do bloco de Atenção Multi-Cabeça, que executa em paralelo múltiplas atenções do tipo Produto Escalar Escalonado, cuja estrutura interna é detalhada em (c). Adaptado de (Vaswani et al., 2017) 15

3.1 Etapas iniciais da metodologia proposta, desde a coleta dos dados genômicos no repositório GISAID até a remoção de *outliers* associados às substituições de aminoácidos, resultando na formação do *Dataset-1* . . . 30
3.2 Distribuição do número de *AASubstitutions* por amostra: (a) antes da remoção de *outliers*, evidenciando alta dispersão, e (b) após a filtragem, com redução da variabilidade e maior homogeneidade do conjunto de dados. 31
3.3 Etapas finais da metodologia, ilustrando a integração dos *embeddings* gerados ao *dataset-1* e a construção do *dataset-3*, utilizado nas análises vetoriais e nos experimentos subsequentes. 34
3.4 Aplicação do método do cotovelo (*Elbow*) para avaliação do número adequado de clusters, com base na variação da soma dos erros quadráticos intra-cluster. 37
3.5 Distribuição das amostras nas duas dimensões resultantes da projeção via t-SNE. A Figura (a) apresenta a distribuição ao longo da primeira dimensão (Dim1), enquanto a Figura (b) corresponde à segunda dimensão (Dim2), evidenciando a estrutura dos dados no espaço reduzido. 40

4.1 Projeção das amostras obtido via t-SNE e agrupamento k-Means. 43
4.2 Mutações mais frequentes associadas a cada cluster obtido com o algoritmo K-Means, destacando padrões mutacionais característico dos agrupamentos identificados. 45
4.3 Projeção das amostras no espaço latente bidimensional obtido via t-SNE. 46
4.4 Heatmap da média de pertinência *fuzzy* de cada variante de preocupação (VOC) em relação aos clusters identificados, indicando o grau de associação entre variantes e agrupamentos. 47
4.5 Análise da ambiguidade *fuzzy* nas amostras, evidenciando perfis mutacionais com maior sobreposição entre clusters e regiões de transição no espaço vetorial. 48
4.6 Proporção *fuzzy* das vinte mutações mais exclusivas para cada *cluster*. Cada barra representa a agregação ponderada da pertinência *fuzzy* das amostras em que a mutação ocorre, indicando seu grau de especificidade em relação aos demais agrupamentos. 49
4.7 Matriz de confusão percentual para o classificador SVM com kernel radial, obtida durante a validação cruzada estratificada (5-fold), ilustrando o desempenho do modelo na distinção entre classes. 54
4.8 Matriz de confusão percentual para o classificador XGBoost, obtida durante a validação cruzada estratificada (5-fold), evidenciando a elevada taxa de acerto e a baixa confusão entre classes. 54
4.9 Matriz de confusão percentual para o classificador Random Forest,obtida durante a validação cruzada estratficada (5-fold), destacando o desempenho do modelo na classificação das variantes. 58
4.10 Matriz de confusão percentual para o classificador k-NN, obtida durante a validação cruzada estratificada (5-fold), apresentando a distribuição dos acertos e erros entre as classes avaliadas. 58

Lista de Tabelas

3.1 Exemplo de amostras da variante Omicron extraídas do campo *AASubstitutions*. 32
3.2 Exemplo de amostras da variante Delta extraídas do campo *AASubstitutions*. 33
3.3 Comparação entre duas amostras semelhantes, com as substituições de aminoácidos distintas realçadas em vermelho. 36
3.4 Avaliação do número de *clusters* com base em métricas internas de qualidade do `k-means`, sendo os valores em vermelho aqueles que representam o melhor desempenho em cada métrica. 36
3.5 Avaliação do número de *clusters* com base em métricas internas de qualidade do `Fuzzy C-means`, sendo os valores em vermelho aqueles que representam o melhor desempenho em cada métrica. 38
3.6 Parâmetros utilizados no algoritmo t-SNE. 39

4.1 Distribuição de variantes VOC por cluster. 44
4.2 Resultados da validação cruzada para o classificador SVM com kernel RBF. 51
4.3 Melhores configurações do XGBoost e suas métricas de desempenho. . . 52
4.4 Resultados da validação cruzada para o classificador Random Forest. . . . 53
4.5 Resultados da validação cruzada para o classificador k-NN. 55
4.6 Comparativo entre os classificadores com base na validação cruzada (5-fold). 56
4.7 Acurácia por modelo e por VOC no conjunto de teste externo. 57

Lista de Acrônimos e Abreviações

ACE2	*Angiotensin-Converting Enzyme 2*
BERT	*Bidirectional Encoder Representations from Transformers*
CNN	*Convolutional Neural Network*
COVID-19	*Coronavirus Disease 2019*
DTM	*Dynamic Topic Modeling*
ELMO	*Embeddings from Language Models*
FCM	*Fuzzy C-Means*
FNN	*Feedforward Neural Network*
FPC	*Fuzzy Partition Coefficient*
GISAID	*Global Initiative on Sharing All Influenza Data*
GLOVE	*Global Vectors for Word Representation*
IA	Inteligencia Artificial
LLM	*Large Language Model*
LSTM	*Long short-Term Memory*
ML	*Machine Learning*
NER	*Named Entity Recognition*
NSPS	*Non-Structural Proteins*
NTD	*N-Terminal Domain*
OMS	Organização Mundial da Saúde
PCA	*Principal Component Analysis*
PE	*Partition Entropy*
PLN	Processamento de Linguagem Natural

RAG	*Retrieval-Augmented Generation*
RBD	*Receptor-Binding Domain*
RF	*Random Forest*
RNN	*Recurrent Neural Network*
S2STM	*Sequence-to-Sequence Transformer Model*
SARS-CoV-2	*Severe Acute Respiratory Syndrome Coronavirus 2*
SSE	*Sum Squared Errors*
SVM	*Support Vector Machine*
t-SNE	*t-distributed Stochastic Neighbor Embedding*
TinyML	*Tiny Machine Learning*
TWE	*Temporal Word Embeddings*
VOC	*Variant of Concern*
VOI	*Variant of Interest*
VUM	*Vaiant Under Monitoring*
XB	*Xie-Beni*
XGBoost	*Extreme Gradient Boosting*

Capítulo 1

Introdução

Este capítulo apresenta uma visão integrada sobre o contexto da pandemia de COVID-19 e sua relação com a necessidade de soluções computacionais avançadas para o monitoramento genômico. Inicialmente, discute-se o cenário global desencadeado pela disseminação do SARS-CoV-2, destacando os desafios impostos pelo surgimento contínuo de novas variantes e os impactos epidemiológicos associados. Na sequência, o texto aborda as principais contribuições da literatura no campo da inteligência artificial aplicada à análise de sequências virais, com ênfase nas estratégias baseadas em aprendizado profundo e PLN, que vêm sendo utilizadas para representar e interpretar dados genômicos de forma mais eficiente e escalável. Dentro dessa perspectiva, são exploradas as motivações que fundamentam a escolha da abordagem metodológica adotada neste trabalho. O capítulo também aprofunda a descrição da estrutura genômica do SARS-CoV-2, abordando suas regiões codificantes e proteínas-chave, com o intuito de fornecer o embasamento biológico necessário para a compreensão das análises que serão desenvolvidas nos capítulos subsequentes.

1.1 A Pandemia de COVID-19: Panorama Global e Impactos

A pandemia de COVID-19, causada pelo vírus SARS-CoV-2, constituiu um dos eventos sanitários, sociais e econômicos mais relevantes do século XXI. O surto teve início em dezembro de 2019, na cidade de Wuhan, província de Hubei, na China, quando a OMS foi alertada sobre casos de pneumonia de causa desconhecida. Em janeiro de 2020, identificou-se o vírus responsável como um novo tipo de coronavírus, denominado SARS-CoV-2, e a doença associada foi oficialmente nomeada de COVID-19 em fevereiro do mesmo ano. Em 11 de março de 2020, a OMS declarou estado de pandemia.

Desde então, mais de 777 milhões de casos foram confirmados em todo o mundo, com mais de 7 milhões de óbitos atribuídos à doença até o primeiro trimestre de 2025. O Brasil registrou mais de 37 milhões de casos e aproximadamente 703 mil mortes, com impactos significativos sobre o sistema de saúde e a organização social (World Health Organization, 2025).

Os efeitos da pandemia transcenderam a saúde pública, afetando diretamente os sis-

temas de saúde, que enfrentaram colapsos hospitalares, escassez de insumos e sobrecarga de profissionais (Ranzani et al., 2021). No plano socioeconômico, as medidas de contenção, como quarentenas e fechamento de fronteiras, desencadearam crises econômicas, aumento do desemprego, instabilidade comercial e retração do produto interno bruto (PIB) em vários países (Deb et al., 2022).

O caráter dinâmico do SARS-CoV-2, com sua alta taxa de transmissão e capacidade de acumular mutações, resultou no surgimento de múltiplas variantes, algumas com potencial aumentado de transmissibilidade, escape imune e gravidade (Scarpa et al., 2024). Esse contexto evidenciou a necessidade de ferramentas computacionais escaláveis, capazes de acompanhar em tempo real as mudanças genômicas e suas implicações epidemiológicas. Dessa forma, a análise automatizada do genoma viral, associada à integração com metadados clínicos e geográficos, surge como uma abordagem promissora no apoio à vigilância genômica.

Com o avanço da tecnologia, especialmente na área de Inteligência Artificial (IA), novas possibilidades surgiram para a pesquisa e compreensão da evolução da COVID-19. Técnicas de Aprendizado de Máquina (ML) permitiram processar e interpretar dados complexos de forma rápida e precisa, contribuindo para a identificação de padrões e correlações que seriam difíceis de detectar por métodos tradicionais. O uso de IA na bioinformática tem se mostrado promissor em diversas frentes, como a previsão da disseminação do vírus, a identificação de variantes emergentes e o desenvolvimento de vacinas (Wang et al., 2023).

A capacidade dessas tecnologias de aprender com dados históricos e em tempo real possibilita a modelagem da evolução viral e a previsão de mutações que podem aumentar a transmissibilidade ou a gravidade da doença. Modelos baseados em algoritmos de aprendizado profundo, como redes neurais e *Transformers*, têm sido empregados para analisar sequências genômicas do SARS-CoV-2. Esses modelos auxiliam na interpretação dos impactos dessas mutações sobre a estrutura e a função das proteínas virais, contribuindo para o entendimento da patogênese do vírus (Lamb et al., 2024).

Entretanto, o uso de IA também apresenta desafios, como a necessidade de grandes volumes de dados de alta qualidade, o poder computacional necessário para processar essas informações e a interpretação dos resultados gerados pelos modelos. Além disso, a natureza dinâmica da pandemia e a constante evolução do vírus exigem atualizações frequentes nos modelos e abordagens adaptativas para manter a precisão e a relevância das análises (Li et al., 2024).

1.2 Estrutura do SARS-CoV-2 e Mecanismo de Infecção

O SARS-CoV-2 é um vírus envelopado, de formato aproximadamente esférico, com diâmetro estimado entre 80 e 220 nanômetros, conforme observado por microscopia eletrônica. Essa estrutura tridimensional é representada na Figura 1.1, que destaca seus principais componentes morfológicos. O vírus possui genoma de RNA de fita simples e polaridade positiva ((+)ssRNA), com cerca de 30 mil nucleotídeos, pertencente ao gênero Betacoronavirus da família Coronaviridae. Sua estrutura é composta por um envelope lipídico derivado da célula hospedeira, que abriga quatro proteínas estruturais principais:

`spike` (*S*), `envelope` (*E*), `membrana` (*M*) e `nucleocapsídeo` (*N*) (Ali, 2024, Andersen et al., 2020).

A proteína `S` atua na interação com o receptor celular *ACE*2 e é fundamental para a entrada viral. Funcionalmente, a proteína `Spike` é dividida em duas subunidades principais, *S*1 e *S*2. A subunidade *S*1 contém o domínio N-terminal (NTD) e o domínio de ligação ao receptor (RBD), sendo este último responsável pelo reconhecimento e ligação ao *ACE*2. Já a subunidade *S*2 inclui elementos estruturais importantes para a fusão da membrana viral com a célula hospedeira, como o peptídeo de fusão (FP), as regiões heptarrepetidas 1 e 2 (HR1 e HR2), além do domínio transmembrana (TM). A clivagem proteolítica entre *S*1 e *S*2 ocorre em um sítio específico (*S*1/*S*2), etapa essencial para ativação da capacidade de fusão da proteína. As proteínas `M` e `E` estão envolvidas na montagem da partícula viral e na estabilidade do vírion, ou seja, do vírus completo e funcional fora da célula hospedeira. A proteína *N* associa-se ao RNA genômico formando o nucleocapsídeo (Fang et al., 2023). A Figura 1.2 apresenta de forma esquemática esses componentes.

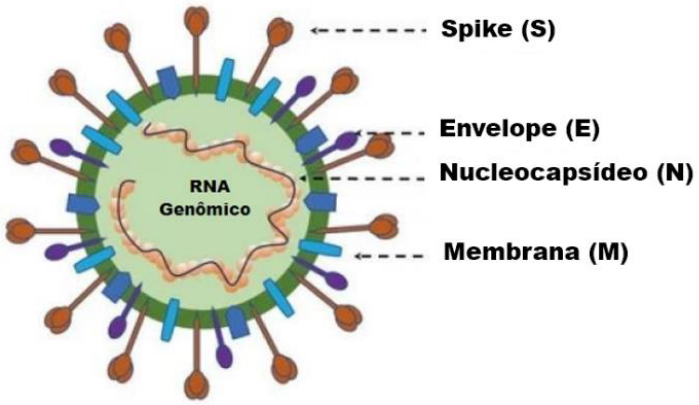


Figura 1.1: Representação estrutural do SARS-CoV-2, destacando seus principais componentes virais e proteínas estruturais. Adaptada de (Ali, 2024)

Além das proteínas estruturais, o genoma viral codifica 16 proteínas não estruturais (nsps), resultantes da clivagem das poliproteínas $pp1a$ e $pp1ab$ traduzidas a partir das regiões $ORF1a$ e $ORF1b$. Essas *nsps* desempenham funções essenciais na replicação e transcrição viral, incluindo atividades de protease (nsp3, nsp5), polimerase (nsp12), helicase (nsp13) e modificação do RNA (nsp16), conforme descrito por (Bai et al., 2022). O vírus também expressa nove proteínas acessórias, como $ORF3a$, $ORF6$, e $ORF8$, que estão envolvidas em mecanismos de evasão imune e regulação da resposta do hospedeiro.

O processo de infecção viral inicia-se com a ligação da subunidade *S*1 da proteína `Spike` ao receptor *ACE*2, presente em células do trato respiratório. Essa interação é facilitada pela protease *TMPRSS*2, que cliva a região *S*2, permitindo a fusão das membranas viral e celular (Hoffmann et al., 2020). Uma vez dentro da célula hospedeira, o RNA viral é traduzido em proteínas virais e replicado com auxílio do complexo de replicação formado pelas *nsps*. A Figura 1.3 ilustra o fluxo desse processo de infecção viral.

Desde seu surgimento no final de 2019, o SARS-CoV-2 tem demonstrado uma notável capacidade de evolução genômica, impulsionada principalmente pela sua natureza como

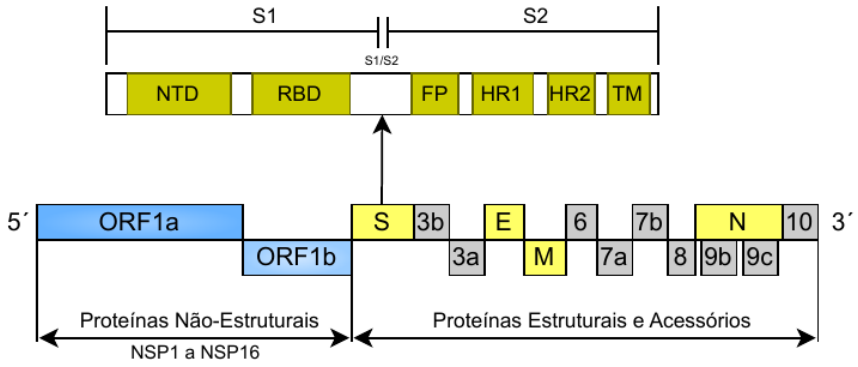


Figura 1.2: Estrutura do genoma do SARS-CoV-2, evidenciando a organização das regiões codificadoras e não codificadoras, bem como as principais proteínas virais associadas. Adaptada de (Fang et al., 2023)

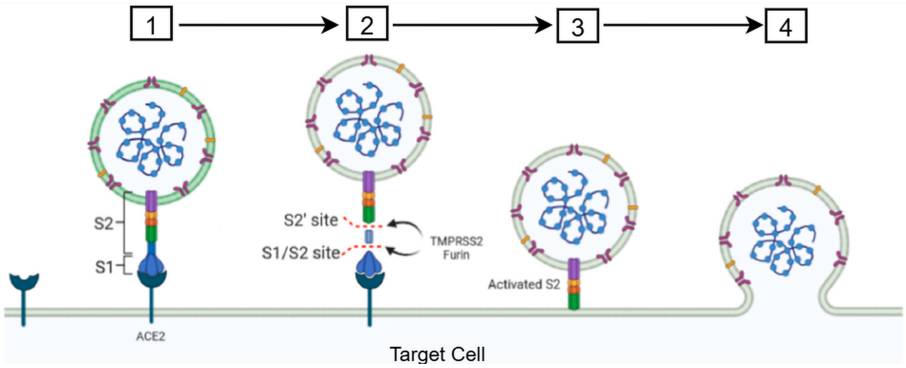


Figura 1.3: Diagrama esquemático da interação entre a proteína Spike do SARS-CoV-2 e o receptor ACE2 da célula hospedeira, ilustrando o mecanismo molecular de entrada viral. Adaptada de (Villoutreix et al., 2021)

vírus de RNA de fita simples positiva. Esse tipo de genoma, ao contrário do DNA, necessita de mecanismos eficazes de correção de erros durante a replicação, o que resulta em uma taxa de mutação relativamente elevada (Lauring e Hodcroft, 2021). A cada ciclo de replicação viral, pequenas alterações, substituições de nucleotídeos, inserções, deleções podem ser acumuladas, especialmente em regiões genômicas com papel funcional crítico, como o gene `Spike`, responsável pela interação com receptores celulares.

Embora muitas dessas mutações sejam neutras, uma fração delas pode conferir vantagens seletivas ao vírus, como maior afinidade ao receptor humano *ACE*2, escape imune parcial ou aumento da transmissibilidade. O acúmulo dessas mutações ao longo do tempo e sua fixação em determinadas linhagens virais conduziram ao surgimento das VOCs, classificadas pela OMS com base em critérios epidemiológicos, clínicos e laboratoriais (World Health Organization, 2025).

As VOCs mais relevantes identificadas até o momento incluem Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2) e Omicron (B.1.1.529), cada uma caracteri-

zada por um conjunto distinto de mutações, sobretudo na proteína `Spike`. Por exemplo, a mutação `N501Y` está presente nas variantes Alpha, Beta e Gamma, estando associada ao aumento da afinidade pelo receptor *ACE*2 (Liu et al., 2022), já a mutação *E*484*K* está fortemente relacionada ao escape de anticorpos neutralizantes, comprometendo parcialmente a eficácia de vacinas (Wang et al., 2021).

O processo evolutivo contínuo do SARS-CoV-2 exige a adoção de estratégias de vigilância genômica em tempo real. A detecção precoce de novas variantes é essencial para orientar decisões relacionadas à saúde pública, tais como restrições de mobilidade, reforço da testagem, revisão da eficácia vacinal e preparação de protocolos clínicos diferenciados. Para isso, consórcios internacionais, como o *Global Initiative on Sharing All Influenza Data* (GISAID), têm desempenhado papel fundamental ao reunir, curar e disponibilizar milhões de genomas virais sequenciados globalmente (Khare et al., 2021).

Entretanto, a escala e a complexidade dos dados genômicos disponíveis impõem desafios computacionais consideráveis. O uso de abordagens tradicionais, como alinhamentos múltiplos de sequências (MSA), embora eficazes para estudos filogenéticos, tornam-se pouco viáveis em bases com milhões de entradas. Assim, cresce o interesse por representações alternativas que abstraiam o conteúdo informativo dos genomas virais de forma compacta, interpretável e compatível com métodos de aprendizado de máquina.

1.3 Estado da Arte

Desde o surgimento do SARS-CoV-2, o acompanhamento de suas variantes tornou-se um dos principais desafios da vigilância epidemiológica global. A rápida acumulação de mutações no genoma viral, especialmente na proteína `Spike`, impacta diretamente a transmissibilidade, o escape imunológico e a eficácia das vacinas, tornando essencial o desenvolvimento de métodos precisos para a detecção de variantes de preocupação. Nesse contexto, abordagens computacionais que classificam automaticamente essas variantes com base em seus perfis mutacionais representam ferramentas valiosas para subsidiar ações de saúde pública.

Pesquisas iniciais se concentraram em abordagens que, embora eficazes, frequentemente dependem de infraestrutura laboratorial. Propostas baseadas em espectroscopia *Raman* com realce de superfície (SERS), por exemplo, alcançaram 100% de acurácia na distinção entre as variantes Beta, Delta e Omicron em amostras de saliva, mas requerem equipamentos especializados (Qin et al., 2024). De forma semelhante, a combinação de qPCR de alta resolução (HRM) com aprendizado de máquina foi apresentada como uma alternativa econômica ao sequenciamento genético, atingindo 95,2% de sensibilidade em amostras clínicas (Promja et al., 2023). Outra inovação notável foi o desenvolvimento de uma plataforma de diagnóstico portátil, baseada em sensores eletroquímicos com nanopartículas de ouro e grafeno, que utiliza *TinyML* embarcado para identificar as variantes Alfa, Beta e Delta com 99,37% de acurácia em menos de um minuto, diretamente de *swabs* (Beduk et al., 2022). Em outra frente, investigou-se a análise de sintomas clínicos para mapear a variante causadora e prever o desfecho da doença, uma estratégia que depende da disponibilidade e qualidade dos registros médicos (Fatima e Ahmad, 2024).

Paralelamente, uma quantidade significativa da pesquisa voltou-se para modelos computacionais supervisionados. Arquiteturas de aprendizado profundo, como o modelo de múltiplos *Transformers* desenvolvido por Qayyum et al. (2024) e as redes híbridas com mecanismo de atenção de Mutawa (2025), foram aplicadas com sucesso para classificar sequências genômicas a partir de dados de nucleotídeos, alcançando acurácias superiores a 99% na distinção entre o SARS-CoV-2 e outros vírus respiratórios. O uso de Redes Neurais Convolucionais (CNNs) também se mostrou eficaz para essa finalidade (Câmara et al., 2022, Azevedo et al., 2024, de Souza et al., 2023, Coutinho et al., 2023). No entanto, o foco principal dessas abordagens reside na distinção entre diferentes vírus, não explorando a diversidade e a dinâmica evolutiva dentro do próprio SARS-CoV-2.

Outros trabalhos focaram na predição de atributos específicos, limitando-se frequentemente à proteína `Spike`. O modelo TEMPO, por exemplo, utilizou *Transformers* para prever futuras mutações pontuais (Zhou et al., 2023), enquanto outras redes neurais foram aplicadas para estimar a severidade clínica da doença (Sokhansanj et al., 2022). Embora o avanço preditivo seja notável, o escopo restrito à proteína `Spike` ignora o panorama global de mutações e as possíveis interações funcionais com outras proteínas. Tentando superar essa visão, um modelo *Sequence-to-Sequence Transformer* foi empregado para investigar a coocorrência de mutações entre as proteínas `Spike` e `M`, um passo importante no reconhecimento das interações genômicas (Ye et al., 2024). Contudo, a análise ainda se manteve focada em um subconjunto pré-definido de genes.

Para uma visão mais holística, pesquisadores começaram a aplicar técnicas de PLN, tratando genomas como “documentos"e mutações como “palavras". Nessa linha, o uso de *Dynamic Topic Modeling* (DTM) e *Temporal Word Embeddings* (TWE) permitiu rastrear assinaturas mutacionais e sua evolução espaço-temporal, associando-as a linhagens e regiões geográficas (Singh et al., 2022). De forma semelhante, modelos de linguagem não supervisionados foram aplicados à vigilância genômica, conseguindo antecipar surtos de casos por meio da análise de entropia (Nagpal et al., 2022). Outras propostas computacionais incluíram o uso de `k-mers`, isto é, subsequências de tamanho fixo extraídas de sequências biológicas, para representar sequências da proteína `Spike`, atingindo alta acurácia na classificação mesmo com apenas 1% dos dados para treinamento (Ali et al., 2021), e a introdução de modelos de aprendizado federado, combinando algoritmos locais como `Random Forest` e `XGBoost` com uma rede neural global, para preservar a privacidade dos dados entre instituições, com acurácia superior a 93% (Chourasia et al., 2023).

Apesar do avanço que representam, essas abordagens ainda apresentam lacunas. Alguns métodos de PLN possuem baixa granularidade na representação das mutações e utilizam uma segmentação temporal rígida (Singh et al., 2022), enquanto outros, embora inovadores para a vigilância, não se aprofundam na caracterização da estrutura latente dos agrupamentos mutacionais (Nagpal et al., 2022). Permanece, portanto, a necessidade de um método que não apenas represente as mutações de forma granular, mas que também explore a topologia dos *clusters* de variantes, identificando perfis de mutação e amostras ambíguas.

Para abordar essas questões, o presente trabalho propõe uma abordagem inovadora, puramente computacional, que integra métodos não supervisionados e supervisionados.

A principal vantagem metodológica reside na utilização de substituições de aminoácidos como unidade principal de informação. Para processar essas substituições, foi empregado o modelo pré-treinado `all-MiniLM-L6-v2`, pertencente à família *Sentence Transformers*. Esse modelo segue a arquitetura *Transformer*, que utiliza camadas de autoatenção (*self-attention*) para atribuir diferentes pesos às relações entre cada substituição na sequência, capturando dependências de curto e longo alcance no perfil mutacional. Inicialmente, as substituições são tokenizadas e mapeadas para vetores iniciais, *embeddings* de entrada, que representam semanticamente cada *token*. Em seguida, as camadas de atenção e *feed-forward* refinam essas representações com base no contexto global da sequência. Por fim, é gerado um vetor denso de 384 dimensões que sintetiza, de forma contextualizada, todo o conjunto de mutações de uma amostra, constituindo a representação vetorial utilizada nas etapas subsequentes de análise. Essa estratégia de vetorização permite preservar relações complexas entre mutações, superando limitações de métodos que operam no nível de nucleotídeos ou que se restringem a genes específicos.

Os experimentos realizados evidenciam a eficácia da proposta em ambas as frentes metodológicas. No contexto supervisionado, os classificadores treinados com esses *embeddings* demonstraram desempenho excepcional, com destaque para o modelo `XGBoost`, que alcançou acurácia de 99,83% e logLoss de apenas 0,0068, mantendo alta robustez mesmo em dados não vistos durante o treinamento. Já na vertente não supervisionada, foram aplicados os algoritmos e `K-Means` e `Fuzzy C-Means`, aliados à projeção bidimensional dos *embeddings* por meio do `t-SNE`, o que permitiu identificar agrupamentos coerentes com as variantes de preocupação, além de revelar assinaturas mutacionais distintas, amostras ambíguas e zonas de transição entre variantes próximas evolutivamente. A análise *fuzzy* possibilitou ainda a caracterização de mutações exclusivas por *cluster* e a quantificação da incerteza na atribuição das amostras, oferecendo uma perspectiva mais refinada sobre a estrutura latente da diversidade viral.

A solução é intrinsecamente escalável, de baixo custo do ponto de vista computacional, uma vez que se baseia em representações vetoriais compactas e em modelos pré-treinados, demandando recursos computacionais moderados mesmo para grandes volumes de dados. Além disso, o método dispensa a necessidade de alinhamento genômico, oferecendo uma abordagem automatizável e original para a vigilância genômica e o estudo da dinâmica evolutiva de patógenos. Ao fornecer uma análise baseada exclusivamente em dados genômicos, o método proposto complementa a análise clínica tradicional, ampliando as perspectivas na detecção e monitoramento de variantes.

1.4 Objetivos da Pesquisa

Com o avanço das estratégias computacionais para vigilância genômica e a crescente disponibilidade de dados mutacionais em escala global, torna-se necessário investigar abordagens alternativas que permitam representar e interpretar a diversidade viral de forma eficiente e contextualizada. No contexto da pandemia de COVID-19, as substituições de aminoácidos observadas no genoma do SARS-CoV-2 constituem uma fonte valiosa de informação, ainda pouco explorada por métodos de aprendizado profundo.

Diante desse cenário, formula-se a seguinte hipótese de pesquisa: A representação

vetorial de substituições de aminoácidos, gerada por modelos de linguagem baseados em *Transformers*, possibilita a captura de relações contextuais e estruturais relevantes entre mutações virais. Quando explorada por meio de abordagens baseadas em representação vetorial contextual e aprendizado profundo, essa estratégia permite identificar padrões latentes nos perfis genômicos, bem como realizar a classificação automatizada de variantes do SARS-CoV-2. Trata-se, portanto, de uma alternativa viável e escalável às metodologias genômicas tradicionais, com potencial para aprimorar a discriminação e contextualização de grupos virais em cenários de vigilância genômica.

Assim, os objetivos deste trabalho foram definidos de modo a estruturar uma metodologia que explore, por meio de representações vetoriais e técnicas de aprendizado de máquina, a informação contida nas substituições de aminoácidos para fins de agrupamento e classificação de variantes virais.

1.4.1 Objetivo Geral

Propor uma metodologia baseada em aprendizado profundo e *embeddings* para representar, agrupar e classificar variantes do SARS-CoV-2, utilizando exclusivamente as substituições de aminoácidos, sem necessidade de alinhamento genômico ou engenharia manual de atributos.

1.4.2 Objetivos Específicos

1. Gerar representações vetoriais a partir de substituições de aminoácidos, utilizando modelos baseados em linguagem natural;
2. Aplicar técnicas de clusterização para investigar a estrutura latente dos perfis mutacionais;
3. Desenvolver modelos de classificação capazes de discriminar variantes virais com base nas representações vetoriais;
4. Avaliar o desempenho dos modelos preditivos com base em métricas de acurácia, equilíbrio entre classes e confiabilidade probabilística;
5. Interpretar os agrupamentos e as classificações a partir das substituições de aminoácidos mais informativas, buscando identificar padrões mutacionais relevantes.

1.5 Organização do Trabalho

Este documento está estruturado em cinco capítulos principais que se complementam de forma progressiva. O Capítulo 1 introduz o tema, contextualizando a relevância da vigilância genômica do SARS-CoV-2 e destacando o papel das técnicas de Inteligência Artificial na análise de substituições de aminoácidos, além de apresentar as motivações, os objetivos do estudo e as lacunas identificadas na literatura. Em seguida, o Capítulo 2 oferece a fundamentação teórica necessária, abordando conceitos centrais de bioinformática genômica, técnicas de representação vetorial com *embeddings*, algoritmos de agrupamento, com ênfase em métodos de clusterização e modelos supervisionados, além dos critérios adotados para avaliação dos métodos. O Capítulo 3 descreve detalhadamente a

metodologia e os experimentos conduzidos, incluindo o processo de coleta e preparação dos dados, a geração de representações vetoriais a partir das substituições de aminoácidos, e a aplicação de algoritmos de clusterização e classificação supervisionada. Na sequência, o Capítulo 4 apresenta os principais resultados obtidos e os discute à luz da literatura especializada, com foco nas estruturas latentes reveladas pelos agrupamentos *fuzzy*, na relação entre os *clusters* e variáveis como variante viral, continente e faixa etária, bem como no desempenho dos classificadores supervisionados na identificação das variantes. Por fim, o Capítulo 5 sintetiza as conclusões parciais do trabalho, propõe perspectivas para estudos futuros e apresenta o cronograma de atividades previstas até a conclusão da dissertação.

1.6 Trabalho Aceitos em Congressos

1. Silva, Alessandro S.; Fernandes, Marcelo A. C.; Câmara, Gabriel B. M. "Abordagem Baseada em Embeddings e Transformers para Vigilância Genômica do SARS-CoV-2". XVII Congresso Brasileiro de Inteligência Computacional, CBIC, 2025.

Este trabalho apresenta uma abordagem computacional para vigilância genômica do SARS-CoV-2 baseada em representações vetoriais de substituições de aminoácidos, geradas por modelos *Transformer*. As mutações são tratadas como unidades semânticas, permitindo a conversão de perfis mutacionais em *embeddings* de alta dimensionalidade sem necessidade de alinhamento genômico. A metodologia é avaliada por meio de técnicas de clusterização e projeção em baixa dimensionalidade, demonstrando a capacidade de identificar padrões estruturais associados às principais variantes virais. Os resultados evidenciam o potencial da proposta como ferramenta escalável e automatizável para análise genômica em larga escala.

2. Silva, Alessandro S.; Fernandes, Marcelo A. C.; Câmara, Gabriel B. M. "Agrupamento Fuzzy de Substituições de Aminoácidos com Embeddings para Análise Genômica Avançada do SARS-CoV-2". VII Congresso Brasileiro de Sistemas Fuzzy, CBFS, 2025.

Neste trabalho, é investigada a aplicação de técnicas de clusterização fuzzy sobre *embeddings* derivados de substituições de aminoácidos do SARS-CoV-2. Utilizando o algoritmo Fuzzy C-Means, o estudo explora a pertinência parcial das amostras a múltiplos *clusters*, permitindo identificar zonas de transição, sobreposição entre variantes e perfis mutacionais ambíguos. A abordagem fuzzy fornece uma interpretação mais flexível da estrutura genômica dos dados quando comparada a métodos particionais tradicionais, contribuindo para uma análise mais refinada da diversidade mutacional e da dinâmica evolutiva viral.

3. Silva, Alessandro S.; Fernandes, Marcelo A. C.; Barbosa, Raquel de M.; Azevedo, Karolayne S. "Classificação de Variantes do SARS-CoV-2 a partir de Embeddings de Substituições de Aminoácidos com Aprendizado Supervisionado". XXII Encontro Nacional de Inteligência Artificial e Computacional, ENIAC, 2025.

Este artigo propõe uma abordagem supervisionada para a classificação de variantes do SARS-CoV-2 utilizando *embeddings* gerados a partir de substituições de aminoácidos. Os vetores obtidos alimentam diferentes classificadores clássicos de aprendizado de máquina, incluindo SVM, Random Forest, k-NN e XGBoost, avaliados por validação cruzada e em

conjunto de dados externo. Os resultados demonstram elevado desempenho preditivo, com destaque para o XGBoost, indicando que representações baseadas exclusivamente em perfis mutacionais são suficientes para discriminar variantes com alta precisão, sem a necessidade de alinhamento genômico.

Capítulo 2

Fundamentação Teórica

A genômica computacional tem sido uma ferramenta indispensável no combate a pandemias, especialmente a de COVID-19. A capacidade de processar e analisar vastos volumes de dados de sequenciamento genômico em tempo hábil permitiu o monitoramento da evolução viral e a identificação de variantes. Nesse contexto, técnicas de aprendizado de máquina e, mais recentemente, de PLN, emergiram como abordagens poderosas, tratando o genoma como uma fonte de informação complexa, passível de ser decodificada por algoritmos avançados.

2.1 Fundamentos de PLN e Representações Semânticas

A representação de dados genéticos sob a ótica do PLN tem ganhado espaço na bioinformática, principalmente em tarefas que envolvem mutações, relações funcionais e predição de fenótipos. Inspirada na analogia entre sequências linguísticas e sequências biológicas, essa abordagem considera elementos como códons, aminoácidos ou sequências proteicas como *tokens* de uma linguagem biológica, que podem ser vetorizados e interpretados por modelos baseados em aprendizado profundo. Esta seção detalha os principais componentes dessa abordagem, incluindo a tokenização, os *embeddings* e a arquitetura *Transformer*, núcleo conceitual dos modelos utilizados neste trabalho.

2.1.1 Tokenização: Transformando Sequências em Unidades de Entrada

A tokenização consiste no processo de segmentação de uma sequência textual em unidades discretas chamadas tokens, que podem corresponder a palavras, subpalavras ou caracteres. Esse procedimento é amplamente utilizado em tarefas de PLN, pois permite que os modelos lidem com textos de forma estruturada, viabilizando a análise automática de grandes volumes de dados (Webster e Kit, 1992).

Métodos clássicos, como o *Byte Pair Encoding* (BPE), operam fundindo iterativamente pares de caracteres mais frequentes, de modo a gerar um vocabulário de subpalavras capaz de representar a maior parte das sequências textuais. Embora bastante eficaz, essa abordagem pode gerar fragmentações indesejadas ou perda de coerência semântica em domínios especializados (Wolf et al., 2020). Para superar tais limitações, estratégias

mais recentes de tokenização sensíveis ao contexto foram propostas, permitindo que os limites entre tokens se ajustem dinamicamente de acordo com a estrutura sintática e o significado do texto (Jones e Dixon, 2024).

Em aplicações de modelos de linguagem, cada elemento tokenizado é transformado em uma representação numérica que pode ser processada pelas redes neurais. Essas sequências de *tokens* constituem as entradas do modelo, servindo como base para a construção de *embeddings* e demais camadas de representação. Para compreender de que forma essas entradas textuais são tratadas internamente, a Figura 2.1 ilustra a composição vetorial de cada *token* em arquiteturas *Transformer*.

Em cada posição da sequência, combina-se a informação proveniente de diferentes vetores: os *embeddings* dos *tokens*, que capturam o significado semântico de cada unidade lexical; os *embeddings* de segmento, que indicam a qual sentença ou parte do texto cada *token* pertence; e os *embeddings* de posição, que codificam a ordem sequencial dos elementos. A soma desses componentes gera a entrada final que será processada pelas camadas internas do modelo.

Ainda na Figura 2.1, o símbolo $[CLS]$ corresponde a um *token* especial adicionado no início da sequência, utilizado como representação agregada para tarefas de classificação, enquanto $[SEP]$ indica um *token* especial de separação entre sentenças ou o final da entrada. Os elementos TOK_i representam os *tokens* obtidos após o processo de tokenização. Já o marcador *EA* refere-se aos *embeddings* de segmento atribuídos a cada *token*, e P_i indica os *embeddings* de posição que codificam a ordem de cada *token* na sequência. A soma desses três vetores resulta na entrada final processada pelas camadas internas do modelo.

O objetivo principal dessa composição vetorial é permitir que o modelo compreenda não apenas o conteúdo individual de cada *token*, mas também sua posição relativa na sequência e, quando aplicável, a segmentação lógica de diferentes partes da entrada. Assim, o modelo é capaz de capturar o significado contextual de cada substituição com base nas demais alterações presentes na sequência. Os *embeddings* posicionais, por sua vez, permitem ao modelo inferir dependências relacionais entre *tokens* mesmo em longas cadeias. Isso contribui para que camadas posteriores de atenção possam operar de forma eficaz, associando mutações específicas a seus contextos evolutivos ou funcionais (Devlin et al., 2018).

2.1.2 Embeddings: Representações Vetoriais Densas e Contextuais

Uma vez tokenizada, cada unidade textual precisa ser convertida em uma representação numérica capaz de ser processada por redes neurais. Esse processo é denominado *embedding* e consiste em mapear cada *token* para um vetor em um espaço de alta dimensionalidade. A ideia central é que palavras ou símbolos semanticamente relacionados ocupem regiões próximas nesse espaço, permitindo que relações de similaridade sejam capturadas de forma quantitativa.

Modelos clássicos como `GloVe`, desenvolvidos por Pennington et al. (2014), produzem vetores fixos, ou seja, independentes do contexto em que um termo aparece. Essa limitação faz com que palavras ambíguas recebam a mesma representação em diferentes

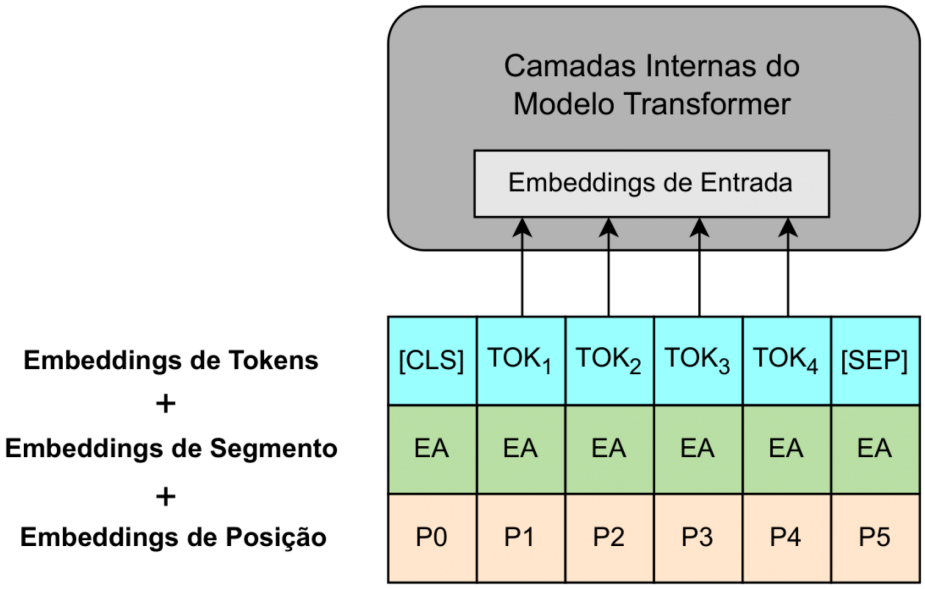


Figura 2.1: Representação vetorial da entrada no BERT, composta pela soma dos *embeddings* de *token*, segmento e posição, que conjuntamente codificam informações semânticas, estruturais e posicionais da sequência. Adaptado de (Devlin et al., 2018)

usos, dificultando a distinção de significados (Pilehvar e Collier, 2016).

Para superar essa limitação, surgiram os *embeddings* contextuais, introduzidos em modelos como `ELMo` de Peters et al. (2018) e o `BERT` proposto por Devlin et al. (2018). Diferentemente dos métodos tradicionais, essas arquiteturas geram representações dinâmicas, em que o vetor atribuído a um *token* depende de toda a sequência em que ele está inserido. Dessa forma, o mesmo termo pode assumir diferentes significados a depender do contexto, aproximando-se mais do funcionamento natural da linguagem.

Esse avanço foi possibilitado principalmente pelo uso de redes recorrentes, no caso do `ELMo`, e pela arquitetura *Transformer*, no caso do `BERT`. Nesses modelos, cada *token* é processado em conjunto com os demais, permitindo capturar relações de curto e longo alcance. Além disso, técnicas de agregação, como a combinação dos vetores ao longo da sequência (*pooling*), possibilitam gerar uma única representação densa para todo o texto, adequada para tarefas de classificação ou análise de similaridade (Reimers e Gurevych, 2019).

Em bioinformática, técnicas baseadas em *embeddings*, originalmente desenvolvidas para representação semântica em PLN (Camacho-Collados e Pilehvar, 2020), têm sido adaptadas com sucesso para modelar entidades como genes, drogas, mutações e interações proteína-proteína. O uso direto de substituições de aminoácidos como *tokens* representa uma inovação metodológica, pois abstrai mutações específicas para uma linguagem textual estruturada, compatível com modelos de PLN.

2.1.3 Arquitetura Transformer: Fundamentos e Mecanismos

A arquitetura *Transformer*, introduzida por Vaswani et al. (2017), representou uma mudança de conceitos nos modelos de linguagem, substituindo redes recorrentes (RNNs) e convolucionais (CNNs) por mecanismos baseados exclusivamente em atenção. A estrutura é modular, composta por blocos empilháveis que operam sobre entradas vetorizadas, possibilitando maior paralelização, escalabilidade e capacidade de aprendizado contextual em longas sequências.

O Núcleo do modelo é o mecanismo de *self-attention*, que permite a cada *token* da sequência comparar-se com todos os demais e ponderar sua relevância contextual. Para isso, cada vetor de entrada é projetado em três representações: *Query* (**Q**), *Key* (**K**) e *Value* (**V**), obtidas por meio de projeção aprendidas.

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V, \tag{2.1}$$

onde $\mathbf{X} \in \mathbb{R}^{n \times d_{\text{model}}}$ é a matriz de entrada, sendo n o número de *tokens* de dimensão d_{model}, e $\mathbf{W^Q}$, $\mathbf{W^K}$, $\mathbf{W^V}$ são matrizes de projeção aprendidas. Essa etapa inicial garante que cada elemento da sequência seja representado em múltiplos espaços de *Query*, *Key* e *Value*, preparando-os para o cálculo de atenção propriamente dito.

A saída da atenção é então calculada por

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2.2}$$

em que o fator $\sqrt{d_k}$ normaliza os valores, evitando saturação da função `softmax` e preservando a estabilidade numérica. Esse mecanismo, denominado *scaled dot-product attention*, constitui a base do *Transformer*, permitindo associar *tokens* relevantes mesmo em posições distantes.

A Figura 2.2 organiza essa arquitetura em três níveis. Em (a) apresenta-se a visão geral, com as camadas de entrada, codificação posicional, blocos do codificador e do decodificador, até a projeção final em probabilidades. Em (b) observa-se a estrutura da atenção multicabeça, e em (c) a formulação interna da atenção por produto escalar escalonado.

Conforme a Figura 2.2(a), o fluxo de dados inicia-se com a camada de *embeddings*, que transforma cada token em um vetor denso de dimensão fixa. Para compensar a ausência de recorrência e convolução, adicionam-se vetores de codificação posicional, geralmente baseados em funções senoidais, que permitem ao modelo distinguir a ordem relativa dos tokens na sequência. Esses vetores são somados aos embeddings e alimentam o primeiro bloco do codificador.

Cada bloco do codificador é composto por uma camada de autoatenção multicabeça seguida de uma rede *feedforward* totalmente conectada, ambas envolvidas por conexões residuais e normalização em camadas (*layer normalization*), o que assegura a estabilidade do treinamento e o fluxo de gradientes. O decodificador adota a mesma estrutura, mas acrescenta uma camada de atenção cruzada que conecta *Queries* oriundos do decodificador com *Keys* e *Values* do codificador. Além disso, aplica-se uma máscara causal na autoatenção, restringindo cada posição a acessar apenas *tokens* anteriores, o que pre-

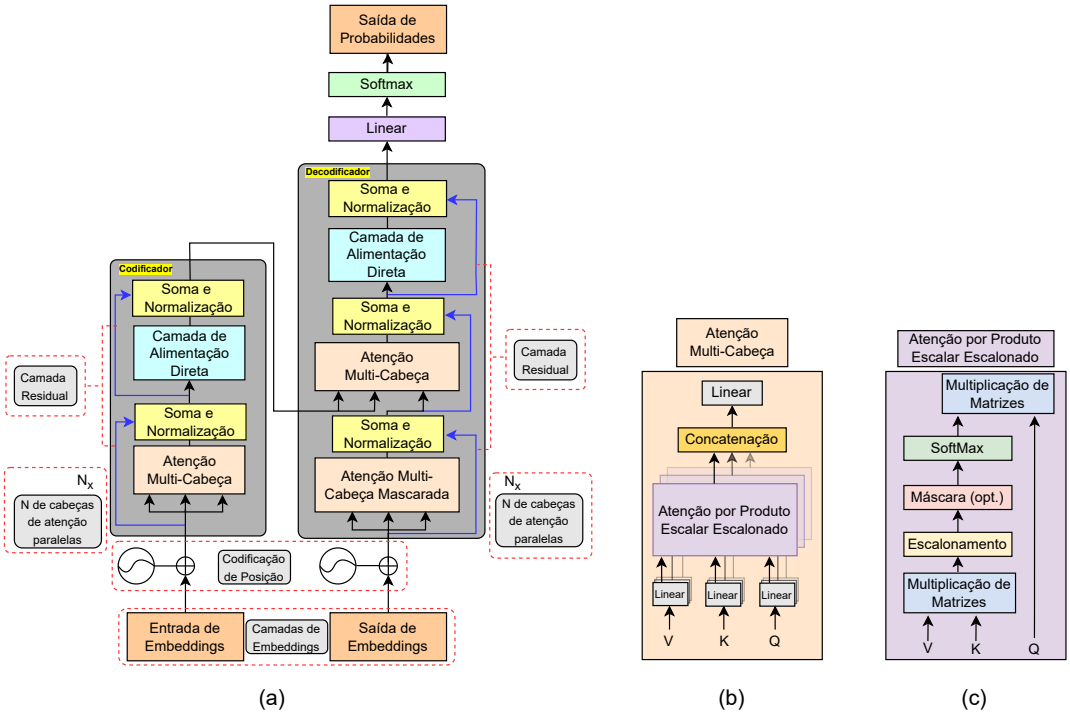


Figura 2.2: A arquitetura geral do modelo *Transformer*, ilustrada em (a), com camadas empilhadas de autoatenção e redes totalmente conectadas. Em (b), apresenta-se o funcionamento do bloco de Atenção Multi-Cabeça, que executa em paralelo múltiplas atenções do tipo Produto Escalar Escalonado, cuja estrutura interna é detalhada em (c). Adaptado de (Vaswani et al., 2017)

serva a natureza autoregressiva da geração. A saída final passa por uma camada linear e uma função *softmax*, que converte os vetores em distribuições de probabilidade sobre o vocabulário.

A Figura 2.2(b) aprofunda o funcionamento da atenção multicabeça. Nela, múltiplos mecanismos de atenção são executados em paralelo, cada qual projetando *queries*, *keys* e *values* em subespaços distintos de representação. As saídas de cada cabeça, calculadas pelo mecanismo da Eq. (2), são então concatenadas e transformadas por uma matriz de projeção, formando a representação final da camada. Essa estratégia possibilita que o modelo capture simultaneamente dependências locais e globais, ampliando sua expressividade.

Por fim, a Figura 2.2(c) apresenta a estrutura interna de uma única cabeça de atenção. Os vetores *query* e *key* são multiplicados para gerar escores de similaridade, escalonados por $\sqrt{d_k}$ e normalizados pela função *softmax*, produzindo pesos de atenção. Esses pesos são aplicados sobre os vetores *value*, resultando em combinações ponderadas que contextualizam cada posição da sequência. Quando aplicado ao decodificador, o mesmo mecanismo incorpora a máscara causal, atribuindo peso nulo a posições futuras e garantindo consistência na geração sequencial.

2.1.4 Modelos Pré-Treinados e Transferência para Domínio Biológico

A evolução dos modelos de linguagem natural foi marcada por um avanço significativo no desempenho a partir da adoção da abordagem de *transfer learning*, no qual grandes modelos são pré-treinados sobre grandes volumes de dados genéricos e, posteriormente, ajustados, ou aplicados diretamente em tarefas específicas. Esse processo ocorre em duas etapas: o pré-treinamento, realizado com objetivos gerais, como a predição de palavras mascaradas utilizando o *Masked Language Modeling* (MLM), e a etapa de adaptação à tarefa-alvo.

Essa adaptação pode ocorrer de diferentes formas, destacando-se principalmente o *fine-tuning*, o *zero-shot learning* e o *few-shot learning*. No *fine-tuning*, os parâmetros do modelo pré-treinado são ajustados utilizando um conjunto de dados rotulado da tarefa de interesse, permitindo especializar suas representações internas para um domínio específico (Devlin et al., 2018). Já no *zero-shot learning*, o modelo é capaz de realizar inferências sobre uma tarefa sem utilizar exemplos rotulados durante a adaptação, explorando apenas o conhecimento semântico adquirido na fase de pré-treinamento (Brown et al., 2020). Por sua vez, o *few-shot learning* caracteriza-se pela adaptação ou avaliação do modelo com base em um número reduzido de exemplos rotulados da tarefa-alvo, explorando a capacidade de generalização do modelo mesmo em cenários de escassez de dados (Brown et al., 2020).

Modelos derivados do `BERT` exemplificam essa tendência, com variantes adaptadas a diferentes contextos. O `BioBERT`, por exemplo, foi treinado sobre textos biomédicos do PubMed e do PMC, mostrando ganhos em tarefas como reconhecimento de entidades nomeadas (NER), perguntas e respostas (QA) e classificação de documentos clínicos (Lee et al., 2019). O `SciBERT`, por sua vez, utilizou um corpus de artigos científicos de múltiplas disciplinas, adotando um vocabulário especializado mais adequado ao jargão científico (Beltagy et al., 2019). Já o `ClinicalBERT` foi ajustado a registros médicos eletrônicos, adaptando-se à linguagem clínica e institucional (Huang et al., 2020). Em outro campo, o `ProtBERT` foi treinado sobre sequências de proteínas da base `UniRef100`, tratando cada aminoácido como um *token*, de modo a capturar relações estruturais e funcionais entre cadeias biológicas (Elnaggar et al., 2022).

Apesar de sua especialização, o emprego direto de modelos como `BioBERT` ou `ProtBERT` sobre representações compactas de mutações pode demandar ajustes adicionais no vocabulário e na forma de entrada. Ainda assim, a literatura tem demonstrado que mesmo modelos generalistas, como o `MiniLM` (Wang et al., 2020), mantêm desempenho satisfatório em tarefas de classificação e similaridade textual, sobretudo devido ao uso de tokenização sublexical do tipo *WordPiece*, que permite decompor termos ausentes em sub-*tokens* conhecidos. Essa característica viabiliza aplicações em domínios técnicos sem necessidade imediata de expansão do vocabulário.

O potencial do uso de modelos pré-treinados no campo biomédico tem sido corroborado por estudos recentes. Cheng et al. (2024) demonstraram que abordagens de *zero-shot learning* podem prever efeitos funcionais de mutações em proteínas sem requerer refinamento adicional. Além disso, arquiteturas multimodais da família `ProtTrans` (Elnaggar

et al., 2022), que combinam representações de sequência e estrutura, evidenciam desempenho competitivo em cenários com dados limitados, explorando o aprendizado autosupervisionado em grandes bases biomoleculares.

Esses avanços reforçam a relevância da transferência de modelos pré-treinados para o domínio biológico, tanto em aplicações de reconhecimento de entidades quanto em tarefas de classificação e análise funcional de mutações. A capacidade de reutilizar representações extraídas de arquiteturas originalmente treinadas em linguagem natural amplia o alcance das técnicas baseadas em *Transformers*, permitindo investigações no contexto genômico e proteico sem depender exclusivamente de grandes volumes de dados rotulados.

2.1.5 Modelo All-MiniLM-L6-v2

O `all-MiniLM-L6-v2` pertence à família dos modelos `MiniLM` e constitui uma versão compacta e eficiente do `BERT`, baseada na arquitetura *Transformer*. Seu design foi otimizado para gerar representações densas em um espaço vetorial de 384 dimensões a partir de sentenças ou fragmentos curtos de texto, oferecendo um equilíbrio entre desempenho e custo computacional. Esse modelo foi pré-treinado em grandes corpora de linguagem geral, incluindo *Wikidata* e *OpenWebText*, o que lhe confere boa capacidade de generalização em tarefas de similaridade semântica e classificação textual, mesmo em cenários com dados rotulados limitados (Wang et al., 2020).

A notação L6 indica a presença de seis camadas *Transformer*, enquanto v2 refere-se à segunda versão do modelo, que incorpora melhorias no processo de treinamento e desempenho em benchmarks de avaliação semântica. O treinamento utilizou o otimizador `AdamW`, amplamente adotado em tarefas de aprendizado profundo por sua eficácia na regularização de pesos e estabilidade durante o ajuste dos parâmetros (Hugging Face, 2024).

Um aspecto central da arquitetura `MiniLM` é a técnica de distilação profunda de autoatenção (*deep self-attention distillation*). Nesse processo, um modelo compacto (*student*) é treinado para imitar diretamente os padrões de atenção e as relações internas aprendidas por um modelo maior (*teacher*, como o `BERT-base`). Diferentemente da distilação tradicional, que exige alinhamento camada a camada, essa estratégia permite maior flexibilidade arquitetural, reduzindo o número de parâmetros e acelerando significativamente o tempo de inferência, ao mesmo tempo em que preserva a competitividade em benchmarks padronizados (Wang et al., 2020).

2.2 Redução de Dimensionalidade em Espaços de Embeddings

A análise de dados em espaços de alta dimensionalidade, como aqueles gerados por *embeddings* de linguagem natural aplicados a sequências biológicas, apresenta desafios conhecidos na literatura como "maldição da dimensionalidade". Esse fenômeno compromete tanto a eficiência computacional quanto a capacidade humana de interpretação visual dos dados, tornando necessária a aplicação de técnicas de redução de dimensionalidade.

De forma geral, a redução de dimensionalidade busca transformar vetores de alta dimensão, definidos em um espaço $\mathbb{R}^d$, em representações mais compactas, geralmente com duas ou três dimensões, preservando, na medida do possível, a estrutura original dos dados. Essa transformação pode ser descrita como uma função $f : \mathbb{R}^d \to \mathbb{R}^k$, onde $k \ll d$.

As técnicas de redução de dimensionalidade podem ser categorizadas em dois grandes grupos, métodos lineares e não lineares. As técnicas lineares assumem que os dados seguem uma estrutura linear subjacente, buscando uma combinação linear das variáveis originais para maximizar a variância ou minimizar a perda de informação. Essas abordagens projetam os dados em um subespaço linearmente relacionado ao espaço original, sendo adequadas para casos onde as relações entre as variáveis são predominantemente lineares (Jolliffe e Cadima, 2016).

Por outro lado, as técnicas não lineares operam sem a suposição de linearidade entre as variáveis. Essas abordagens são projetadas para preservar estruturas de vizinhança local ou global que emergem de relações complexas entre os dados, muitas vezes assumindo que as amostras estão distribuídas sobre uma variedade de baixa dimensionalidade imersa em um espaço de maior dimensão. A utilização de técnicas não lineares é recomendada quando os dados apresentam padrões intrínsecos que não podem ser representados adequadamente por transformações lineares (McInnes et al., 2020).

2.2.1 t-SNE: Teoria e Funcionamento

O *t-Distributed Stochastic Neighbor Embedding* (`t-SNE`), proposto por van der Maaten e Hinton (2008), é uma técnica de redução de dimensionalidade não linear amplamente utilizada para visualização de dados complexos em espaços de duas ou três dimensões. Seu objetivo principal é preservar as relações de vizinhança local entre os pontos, permitindo uma representação gráfica mais intuitiva de agrupamentos e estruturas latentes.

O funcionamento do método baseia-se na conversão das distâncias entre os pontos no espaço original em distribuições de probabilidade que refletem sua similaridade. Cada ponto da base de dados é associado a uma distribuição Gaussiana centrada em si, de modo que a probabilidade de seleção de seus vizinhos diretos seja proporcional à proximidade relativa. Esse processo é ajustado por um parâmetro denominado perplexidade, que controla a densidade efetiva da vizinhança considerada.

As probabilidades locais calculadas no espaço original são então combinadas em uma distribuição conjunta simétrica. No espaço de baixa dimensionalidade, por sua vez, as relações de similaridade entre os pontos são representadas por uma distribuição *t-Student* com um grau de liberdade. Essa escolha é fundamental, pois a distribuição *t-Student* possui caudas mais longas do que a Gaussiana, permitindo atenuar o chamado *crowding problem*, que ocorre quando pontos distantes acabam projetados de forma artificialmente próxima.

A qualidade da projeção é avaliada por meio da divergência de *Kullback–Leibler* entre as distribuições calculadas nos espaços de alta e baixa dimensionalidade. O processo de otimização busca minimizar essa divergência por gradiente descendente, com o auxílio de mecanismos como `momentum` e `early exaggeration`. O `momentum` acumula uma fração da atualização anterior na atualização corrente, suavizando oscilações e acelerando

o avanço ao longo de regiões pouco inclinadas do espaço de parâmetros. Já o `early exaggeration` amplifica temporariamente as probabilidades de vizinhança na fase inicial, ampliando as distâncias entre grupos e facilitando a separação preliminar antes do refinamento final. Em conjunto, esses termos, além da taxa de aprendizado, influenciam a velocidade de convergência e a qualidade da organização local preservada pelo *t-SNE*.

Embora o *t-SNE* seja bastante eficaz na preservação de estruturas locais e na revelação de agrupamentos visuais em *embeddings*, ele também apresenta limitações conhecidas. Entre elas, destacam-se a forte dependência de hiperparâmetros como perplexidade e taxa de aprendizado, a variabilidade dos resultados entre execuções diferentes e a ausência de um mapeamento explícito que permita projetar diretamente novos pontos sem repetir todo o processo de otimização. Ainda assim, sua aplicação em análises exploratórias tem se mostrado valiosa em diversos contextos científicos, incluindo estudos em bioinformática e aprendizado profundo.

2.3 Técnicas de Clusterização para Análise de Embeddings

O avanço das representações vetoriais de dados biológicos por meio de modelos de linguagem natural, especialmente *embeddings* gerados por *Transformers*, tem motivado o uso crescente de técnicas de aprendizado não supervisionado na análise exploratória de grandes conjuntos genômicos. A clusterização, neste contexto, surge como um método importante para detectar padrões latentes e estruturas subjacentes em espaços de alta dimensionalidade (Zvyagin et al., 2022).

De maneira geral, a clusterização visa agrupar elementos similares com base em alguma medida de proximidade, sem a necessidade de rótulos ou categorias prévias. Quando aplicada a *embeddings* essa técnica permite identificar grupos que compartilham características estruturais ou funcionais, ainda que não estejam explicitamente representadas nos dados originais (Singh et al., 2022).

Diferentes algoritmos de agrupamento foram desenvolvidos para capturar distintos aspectos da estrutura dos dados. Métodos clássicos como o `K-means` assumem a existência de fronteiras bem definidas entre os grupos e atribuem cada elemento exclusivamente a um único *cluster*. De acordo com Sokhansanj et al. (2022), abordagens clássicas como o `K-means`, apesar de amplamente utilizadas, podem não capturar adequadamente a complexidade de dados biomoleculares, especialmente em cenários com padrões sobrepostos ou alta heterogeneidade genética, como os observados nas mutações do SARS-CoV-2. Para contornar essa rigidez, técnicas como o `Fuzzy C-Means` (FCM) introduzem um modelo de associação probabilística, no qual cada amostra pertence parcialmente a múltiplos grupos. Essa característica é valiosa na análise de perfis mutacionais virais, nos quais amostras híbridas ou transitórias entre variantes são comuns (Bezdek, 1981).

A combinação entre *embeddings* e algoritmos de clusterização permite representar relações não triviais em dados genômicos, abrindo caminho para investigações mais interpretáveis da diversidade evolutiva viral. Nesse sentido, a análise de agrupamentos em espaços vetoriais de alta dimensionalidade oferece uma alternativa ao alinhamento genô-

mico tradicional, viabilizando aplicações mais escaláveis e adaptáveis em cenários de vigilância molecular (Nagpal et al., 2022).

2.3.1 K-means Clustering

O `K-means` é um dos algoritmos mais amplamente utilizados para particionamento de dados em *k* grupos, sendo eficaz em contextos de alta dimensionalidade, como a análise de *embeddings*. O objetivo central do `K-means` é minimizar a soma das distâncias quadráticas entre os pontos de dados e seus respectivos centroides, buscando encontrar uma configuração de clusters que minimize a variância intra-cluster (MacQueen, 1967, Lloyd, 1982).

Formalmente, o problema de otimização pode ser descrito como a busca de uma partição que minimize a soma das distâncias entre cada ponto e o centróide de seu grupo, sendo este último definido como a média dos elementos pertencentes ao cluster (Lloyd, 1982).

O algoritmo segue uma abordagem iterativa do tipo *Expectation-Maximization* (EM), que pode ser descrita da seguinte forma: primeiro, são escolhidos K centróides iniciais, de forma aleatória ou utilizando estratégias como o `K-means++`. Em seguida, cada ponto é atribuído ao *cluster* mais próximo com base em uma métrica de distância, geralmente a euclidiana. Após essa etapa, os centróides são recalculados como a média dos pontos atribuídos a cada *cluster*. Esse processo se repete até que as atribuições não sofram mais alterações significativas ou até que a variação dos centróides entre iterações fique abaixo de um limiar predefinido.

Um aspecto importante, e que deve ser levado em consideração na implementação do `k-means`, é que o algoritmo tradicionalmente utiliza a métrica de distância `Euclidiana` para medir a proximidade entre os pontos e os centróides. No entanto, essa escolha pode não ser a mais apropriada em todos os cenários, especialmente quando os dados apresentam correlações não lineares, variância heterogênea entre dimensões ou estruturas intrinsecamente não euclidianas.

Dependendo da natureza dos dados, outras métricas podem ser empregadas, como a distância de `Manhattan` ou L_1, distância de `Mahalanobis`, ou ainda a distância `Cosseno`, muito útil em espaços vetoriais de alta dimensionalidade como os *embeddings*. A escolha da métrica de distância influencia diretamente a forma e a separabilidade dos *clusters*, devendo, portanto, ser cuidadosamente considerada conforme o domínio do problema e as propriedades dos dados em análise (Aggarwal et al., 2001).

Outro desafio fundamental é a definição do número de clusters K. Diversos métodos são empregados para estimar esse valor de forma robusta. O método do cotovelo (*Elbow Method*) avalia a soma das distâncias intra-cluster para diferentes valores de K, identificando o ponto em que a redução dessa medida deixa de ser significativa, formando o "cotovelo" da curva. O índice de silhueta (*Silhouette Score*) mede o grau de coesão e separação entre clusters, variando entre -1 e 1, e assume valores mais altos quando os grupos são mais bem definidos (Rousseeuw, 1987). O índice de Calinski-Harabasz avalia a razão entre a variabilidade entre grupos e a variabilidade interna, sendo valores maiores indicativos de separações mais consistentes (Calinski e Harabasz, 1974). Já o índice de Davies-Bouldin mede a relação entre a dispersão intra-cluster e a separação entre centrói-

des, sendo desejável que seu valor seja o menor possível (Davies e Bouldin, 1979).

A seleção do número ótimo de *clusters* depende da aplicação específica, da natureza dos dados e da métrica adotada. Frequentemente, múltiplos critérios são usados em conjunto para fundamentar a escolha de K de forma robusta.

O `K-means` apresenta simplicidade computacional e boa escalabilidade, porém assume que os *clusters* possuem formas esféricas e número similar de amostras, sendo sensível a *outliers* e inicializações diferentes, além de impor uma atribuição rígida de cada ponto a apenas um *cluster* (Hartigan e Wong, 1979). Essas características podem impactar a análise de *embeddings*, que frequentemente não apresentam distribuições gaussianas ou tamanhos de *cluster* homogêneos.

2.3.2 Fuzzy C-means Clustering

O `Fuzzy C-Means` (FCM) é uma extensão do `K-means` baseada na teoria dos conjuntos *fuzzy*, foi proposta por Dunn (1973) e aprimorada por Bezdek (1981). Diferentemente do `k-means`, que realiza uma atribuição *hard clustering*, o FCM permite que cada amostra possua graus de pertencimento a múltiplos grupos. Essa característica torns o algoritmo útil na análise de dados biológicos, nos quais a ambiguidade e a sobreposição entre classes são fenômenos recorrentes.

No FCM, cada amostra recebe valores de pertinência que indicam o quanto ela está associada a cada *cluster*. Esses valores são regulados por um parâmetro denominado expoente de fuzzificação m, que controla o nível de difusão da partição. Quando $m = 1$, o método se reduz ao particionamento rígido tradicional, enquanto valores maiores de m produzem partições mais difusas. A atualização dos centróides ocorre por meio de médias ponderadas, nas quais os graus de pertinência determinam a influência relativa de cada amostra. Da mesma forma, a atualização dos graus de pertinência é feita de acordo com a distância de cada ponto em relação aos centróides, de modo que pontos mais próximos de um centróide apresentam maior associação a ele.

O algoritmo é iterativo e converge quando os centróides deixam de variar significativamente entre iterações. O resultado é expresso em uma matriz de pertinência, que pode fornecer informações adicionais, como o grau máximo de confiança atribuído a cada amostra, medidas de entropia ou exclusividade fuzzy por atributo. Tais indicadores são úteis, por exemplo, para identificar genes ou mutações com papel discriminante em cada cluster.

A determinação do número ótimo de *clusters* K no contexto *fuzzy* pode ser orientada por métricas específicas de validade de partição. Entre as mais utilizadas, destaca-se o `Coeficiente de Partição Fuzzy` (FPC), que quantifica o grau de nitidez da partição, assumindo valores mais altos quando as pertinências são próximas de 1 (Verma et al., 2023), a `Entropia de Partição` (PE), que mede a incerteza média nas associações *fuzzy*, apresentando valores mais elevados em cenários de maior sobreposição (Verma et al., 2023), e o `Índice de Xie-Beni` (XB), que avalia simultaneamente a compactação interna e a separação entre *clusters*, sendo desejável que seus valores sejam baixos (Xie e Beni, 1991).

Em síntese, o FCM oferece uma alternativa matematicamente sólida e biologicamente

coerente para explorar estruturas complexas em dados de alta dimensionalidade, como os *embeddings* gerados a partir de substituições de aminoácidos.

2.4 Aprendizado Supervisionado

O aprendizado supervisionado é uma abordagem de modelagem estatística e computacional em que um modelo é treinado a partir de um conjunto de exemplos rotulados, com o objetivo de aprender uma função que mapeie entradas para saídas desejadas. Em bioinformática, essa metodologia é amplamente aplicada em tarefas como a classificação de variantes virais, predição de efeitos mutacionais e análise de expressão gênica (Kourou et al., 2015).

De forma geral, essa técnica parte de um conjunto de treinamento composto por pares de entrada e saída conhecidos. A partir desses dados, o modelo busca ajustar seus parâmetros de modo a minimizar uma função de perda, que mede a discrepância entre as predições realizadas e os rótulos reais. O objetivo é que, após o treinamento, o modelo seja capaz de generalizar o conhecimento adquirido para amostras não vistas anteriormente, realizando previsões consistentes em novos cenários.

Na tarefa de classificação de variantes virais, diferentes algoritmos de aprendizado supervisionado podem ser empregados. Cada método apresenta pressupostos próprios, diferentes estratégias de particionamento do espaço de atributos e distintas formas de generalização, o que possibilita comparar abordagens baseadas em margens, árvores de decisão, métodos de *ensemble* e técnicas de vizinhança.

2.4.1 Modelos Supervisionados para Classificação de Variantes

O algoritmo *Support Vector Machine* (`SVM`), proposto por Cortes e Vapnik (1995), é uma técnica de aprendizado supervisionado amplamente reconhecida por sua fundamentação sólida e sua capacidade de generalização em tarefas de classificação. O princípio fundamental do `SVM` consiste em encontrar um hiperplano que separe duas classes de dados com a maior margem possível, ou seja, a maior distância entre os pontos mais próximos das duas classes, conhecidos como vetores de suporte.

Em sua formulação clássica, o `SVM` busca maximizar essa margem, o que equivale a minimizar a norma do vetor de pesos que define o hiperplano separador. A solução ótima é obtida a partir de uma combinação linear de exemplos de treinamento, sendo que apenas os vetores de suporte, isto é, as amostras mais próximas da fronteira, contribuem diretamente para a definição da separação.

Quando os dados não são linearmente separáveis, o modelo é estendido por meio da introdução de variáveis de folga, permitindo violações marginais das condições de separação. Esse mecanismo, conhecido como *soft margin*, incorpora um hiperparâmetro C que controla o equilíbrio entre a maximização da margem e a penalização de erros de classificação.

Além disso, para lidar com problemas não lineares, o `SVM` emprega o chamado truque do *kernel*, que consiste em mapear implicitamente os dados para um espaço de características de dimensão mais elevada, onde se busca um separador linear. Esse procedimento

é realizado por meio de funções de `kernel` que calculam produtos internos sem a necessidade de realizar explicitamente a transformação. Entre os *kernels* mais utilizados destacam-se o polinomial e o de base radial (RBF), este último particularmente adequado para capturar fronteiras complexas e não lineares.

Temos também o algoritmo *k-Nearest Neighbors* (`k-NN`), formalmente descrito por Cover e Hart (1967), é um dos métodos mais simples e intuitivos de aprendizado supervisionado. Trata-se de um classificador baseado em instâncias, cuja decisão é determinada pela proximidade entre amostras no espaço de atributos. Em contraste com modelos paramétricos como `SVM` ou `Random Forest`, o `k-NN` não constrói um modelo explícito durante o treinamento, sendo considerado um método do tipo *lazy learner*.

O funcionamento do `k-NN` consiste em identificar, para uma nova amostra, os k vizinhos mais próximos no conjunto de treinamento, com base em uma métrica de distância predefinida. A classe atribuída à amostra é aquela mais frequente entre seus vizinhos. A distância euclidiana é a métrica mais utilizada, mas alternativas como as distâncias de `Manhattan`, `Minkowski` ou `Cosseno` também são empregadas dependendo das propriedades dos dados.

Um resultado teórico importante estabelecido por Cover e Hart (1967) mostra que, sob certas condições, o classificador `1-NN` possui uma taxa de erro assintótica limitada a, no máximo, o dobro da taxa de erro do classificador Bayesiano ótimo. Essa propriedade destaca a robustez teórica do método e justifica sua ampla adoção em diferentes áreas. A escolha do parâmetro k influencia diretamente o comportamento do algoritmo, valores pequenos tendem a gerar classificadores mais sensíveis ao ruído, enquanto valores grandes suavizam excessivamente as fronteiras de decisão, aumentando o viés.

2.4.2 Modelos Baseados em Ensemble

O algoritmo *Random Forest*, introduzido por Breiman (2001), é uma técnica de aprendizado supervisionado baseada em métodos de *ensemble*, voltada para tarefas de classificação e regressão. Ele se fundamenta na combinação de múltiplas árvores de decisão, de modo que cada árvore atua como um classificador fraco e o resultado final é obtido por votação majoritária. A principal inovação desse método é a incorporação de duas fontes de aleatoriedade: a amostragem com reposição (*bootstrap*), que gera diferentes conjuntos de treinamento para cada árvore, e a seleção aleatória de subconjuntos de atributos em cada divisão de nó. Essa diversidade entre as árvores reduz a correlação entre os classificadores individuais, aumentando a capacidade de generalização e tornando o modelo menos suscetível ao *overfitting*.

Do ponto de vista teórico, o desempenho da floresta depende de dois fatores principais, a força das árvores individuais e a correlação entre elas. A força refere-se à precisão média de cada árvore, enquanto a correlação expressa o grau de similaridade entre seus erros. Uma floresta eficiente deve maximizar a força de cada classificador ao mesmo tempo em que minimiza a correlação, garantindo maior acurácia do conjunto. Breiman (2001) demonstrou que, em certas condições, o erro de generalização do `Random Forest` converge para um valor limite à medida que o número de árvores cresce, reforçando a robustez do método.

Já o algoritmo *eXtreme Gradient Boosting* (`XGBoost`), proposto Chen e Guestrin (2016), representa uma das implementações mais eficientes e confiáveis da técnica de *boosting* de gradiente, especialmente projetada para desempenho computacional em larga escala. Fundamentado no modelo de aprendizado aditivo, o `XGBoost` busca construir um conjunto de preditores, geralmente árvores de decisão, de forma sequencial, onde cada novo preditor é ajustado com base nos resíduos ou erros do conjunto anterior.

A função objetivo do `XGBoost` combina dois termos: uma função de perda, que mede a discrepância entre as predições e os rótulos reais, e um termo de regularização, que penaliza a complexidade do modelo. Essa regularização explícita controla o crescimento das árvores e previne sobreajuste, favorecendo modelos mais simples e com maior capacidade de generalização.

Uma inovação importante do `XGBoost` é o uso simultâneo do gradiente e do hessiano da função de perda. Essa expansão de segunda ordem permite que cada nova árvore seja ajustada levando em conta tanto a inclinação quanto a curvatura local da superfície de erro, resultando em atualizações mais estáveis e precisas em comparação com métodos de *boosting* tradicionais.

Além da base teórica, o algoritmo inclui diversas melhorias práticas: poda de árvores baseada em ganho mínimo, suporte nativo para dados esparsos, paralelização eficiente na construção das árvores e uso de técnicas como taxa de aprendizado para reduzir a influência de cada nova árvore no modelo final. Outro aspecto central é a regularização dual (parâmetros γ e λ), que controla a complexidade estrutural e os pesos das folhas, equilibrando precisão e simplicidade.

Graças a essa combinação de fundamentos teóricos e otimizações práticas, o `XGBoost` se consolidou como uma das ferramentas mais robustas para tarefas de classificação e regressão em diferentes domínios, sendo especialmente eficaz em cenários de grande escala e com dados de alta dimensionalidade.

2.4.3 Considerações sobre Escolha de Algoritmo

A escolha do algoritmo de classificação mais adequado depende de múltiplos fatores, como o tamanho do conjunto de dados, a distribuição das classes, a linearidade da separação entre categorias e o grau de interpretabilidade desejado. Em geral, a `SVM` é recomendada para conjuntos de dados com separabilidade complexa e tamanho moderado, o `Random Forest` para cenários com alta dimensionalidade e interesse em interpretação de *features*, o `XGBoost` para grandes volumes de dados e alto desempenho preditivo e o `k-NN` para problemas onde a relação de proximidade espacial é evidente e os dados são previamente normalizados.

2.5 Métricas de Avaliação

A avaliação de desempenho de modelos de classificação é um aspecto crítico em qualquer tarefa de aprendizado supervisionado, pois influencia diretamente a interpretação dos resultados e a escolha do modelo mais apropriado para um determinado problema. As métricas de avaliação têm origem nas teorias estatísticas de análise de decisões e de teste de

hipóteses, sendo amplamente empregadas em contextos de bioinformática, medicina e ciências de dados (Sokolova e Lapalme, 2009).

2.5.1 Acurácia

A acurácia, definida como a proporção de predições corretas em relação ao total de amostras avaliadas, é uma métrica amplamente utilizada para medir o desempenho de modelos de classificação. Formalmente, ela é expressa como

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3}$$

onde TP e TN representam os verdadeiros positivos e negativos, respectivamente, e FP e FN indicam os falsos positivos e negativos.

Apesar de sua popularidade, a acurácia é uma métrica que apresenta viés significativo em cenários com desbalanceamento de classes, uma vez que é diretamente influenciada pelas prevalências marginais ou frequência das classes, e pelo viés do modelo, ou seja, a tendência a predizer determinadas classes com mais frequência. Como destacado por Powers (2011), essa medida pode mascarar desempenhos inadequados quando o modelo acerta majoritariamente a classe majoritária, mesmo falhando sistematicamente na predição da classe minoritária.

Além disso, a acurácia não distingue a natureza dos erros cometidos, tratando da mesma forma um falso positivo e um falso negativo, o que pode ser inadequado em contextos de bioinformática ou clínicos, onde os custos de diferentes erros são assimétricos. Outro problema fundamental é sua insensibilidade ao acaso, um classificador que realiza predições aleatórias pode atingir níveis elevados de acurácia em conjuntos de dados altamente desbalanceados.

Assim, embora a acurácia permaneça útil como uma medida agregada de desempenho, seu uso isolado pode conduzir a interpretações equivocadas sobre a capacidade discriminativa de um modelo. Por esse motivo, recomenda-se complementar sua análise com métricas que levem em conta o desempenho por classe e os desvios em relação ao acaso.

2.5.2 Logarithmic Loss (LogLoss)

A métrica *Logarithmic Loss*, ou simplesmente *LogLoss*, é amplamente utilizada na avaliação modelos de classificação probabilística. Diferente da acurácia, que considera apenas se a classe prevista corresponde ao rótulo verdadeiro, o *LogLoss* também leva em conta o nível de confiança atribuído à predição. Em outras palavras, a métrica penaliza de forma mais intensa os erros cometidos com altas probabilidades, refletindo a confiabilidade do modelo em suas estimativas.

No caso binário, o *LogLoss* avalia a diferença entre a probabilidade prevista para uma determinada classe e o valor real do rótulo. Quando o modelo atribui uma probabilidade elevada à classe incorreta, a penalidade é muito maior do que quando a probabilidade atribuída é baixa. Esse comportamento demonstra a aversão do *LogLoss* a erros "seguros", em que o modelo está confiante, mas incorreto.

Em cenários multiclasse, a métrica é generalizada de forma que apenas a probabilidade associada à classe verdadeira contribui para o cálculo. Isso é possível graças à codificação *one-hot*, em que cada amostra é representada como um vetor indicando a classe correta. Assim, o *LogLoss* pode ser entendido como a média da entropia cruzada entre a distribuição real dos rótulos e a distribuição de probabilidades produzida pelo modelo.

Essa ligação direta com a entropia cruzada faz do *LogLoss* não apenas uma métrica de avaliação, mas também uma das funções de perda mais utilizadas no treinamento de modelos probabilísticos, especialmente em redes neurais aplicadas a problemas de classificação (Tervent et al., 2025).

2.5.3 F1-Macro

A métrica F_1-Macro é uma generalização da tradicional F_1-score, amplamente utilizada na avaliação de classificadores em problemas de múltiplas classes. Enquanto o *F1-score* clássico combina precisão e revocação em um único valor para cenários binários, o *F1-Macro* estende esse raciocínio ao calcular o *F1-score* de cada classe individualmente e, em seguida, computar a média aritmética desses valores. Dessa forma, todas as classes são tratadas com a mesma importância, independentemente de sua frequência no conjunto de dados, o que é particularmente relevante em contextos de desbalanceamento entre classes.

Para uma classe c_i, os valores de Precisão e Revocação são definidos como

$$\text{Precisão}_{c_i} = \frac{TP_{c_i}}{TP_{c_i} + FP_{c_i}}, \quad \text{Revocação}_{c_i} = \frac{TP_{c_i}}{TP_{c_i} + FN_{c_i}} \tag{2.4}$$

e o F_1-score por classe é a média harmônica entre a precisão e a revocação dada por

$$F_1^{c_i} = \frac{2 \cdot \text{Precisão}_{c_i} \cdot \text{Revocação}_{c_i}}{\text{Precisão}_{c_i} + \text{Revocação}_{c_i}}. \tag{2.5}$$

A média macro do F_1-score, denominada F_1-Macro, é então definida como

$$F_1\text{Macro} = \frac{1}{C} \sum_{i=1}^{C} F_1^{c_i} \tag{2.6}$$

onde C é o número total de classes.

Segundo Sokolova e Lapalme (2009), o F_1-Macro é adequado quando se deseja uma avaliação equilibrada do desempenho em todas as classes, sem favorecimento às classes mais frequentes. Contudo, essa métrica também apresenta limitações. Por ser uma média simples, o F_1-Macro pode ser excessivamente penalizado por desempenhos ruins em classes raras, mesmo quando o desempenho geral do modelo em termos de amostras estáveis é satisfatório.

Outra característica relevante do F_1-Macro é sua invariância frente à variação do número de exemplos verdadeiros negativos (TN), refletindo seu foco exclusivo na capacidade do modelo de identificar corretamente as instâncias positivas e minimizar os erros

de omissão (FN) e comissão (FP). Isso o torna adequado para tarefas como classificação de textos ou padrões biológicos, onde o interesse se concentra na detecção correta das ocorrências relevantes, mesmo na presença de uma classe negativa extensa e heterogênea.

2.5.4 Métricas Complementares de Avaliação

Embora a Acurácia, o LogLoss e o F1-Macro constituam as métricas principais na análise de classificadores, especialmente pela sua capacidade de refletir respectivamente a taxa de acertos globais, a confiabilidade probabilística, essas medidas não são suficientes para capturar todas as nuances do desempenho preditivo em cenários complexos. Por essa razão, elas são complementadas por outras métricas que permitem avaliar aspectos específicos, como a capacidade discriminativa em diferentes limiares, a estabilidade em contextos de desbalanceamento entre classes e a relevância de erros assimétricos em aplicações práticas. As métricas complementares discutidas a seguir oferecem, portanto, uma visão mais abrangente e detalhada do desempenho dos modelos.

A primeira delas é a Área sob a Curva ROC (AUC-ROC), que mede a capacidade discriminativa de um modelo binário ao avaliar a relação entre taxa de verdadeiros positivos e taxa de falsos positivos em diferentes limiares de decisão. Seu valor corresponde à probabilidade de que o modelo atribua uma pontuação maior a uma instância positiva do que a uma negativa escolhida ao acaso, sendo equivalente ao teste de Wilcoxon-Mann-Whitney (Hanley e McNeil, 1982). Uma vantagem dessa métrica é sua independência da distribuição entre classes, o que a torna adequada para cenários desbalanceados (Fawcett, 2006). No entanto, ela pode ser enganosa quando apenas uma faixa específica da curva é relevante ou quando curvas distintas geram a mesma área total.

A Sensibilidade Média, também chamada de *macro recall*, calcula a proporção média de instâncias corretamente identificadas como positivas em cada classe. Ela trata todas as classes igualmente, independentemente de sua frequência, o que a torna estável em cenários desbalanceados (Opitz, 2024). Já a Especificidade Média avalia a proporção de instâncias negativas corretamente classificadas em cada classe, sendo essencial em contextos nos quais falsos positivos implicam custos elevados, como em aplicações clínicas (Grandini et al., 2020).

Por fim, a Acurácia Balanceada corresponde à média do *recall* entre todas as classes, oferecendo uma visão menos enviesada do desempenho em cenários desproporcionais. Diferentemente da acurácia tradicional, ela evita que classes majoritárias dominem a avaliação, sendo especialmente adequada em contextos biomédicos ou de detecção em que classes minoritárias têm maior relevância (Farhadpour et al., 2024).

Em conjunto, essas métricas ampliam a interpretação dos resultados, permitindo identificar pontos fortes e limitações dos modelos sob diferentes perspectivas. Essa diversidade de critérios contribui para análises mais consistentes, sobretudo em aplicações biomédicas, nas quais a confiabilidade das previsões é tão relevante quanto a sua exatidão.

Capítulo 3

Proposta, Métodos e Experimentos

Este capítulo apresenta a proposta, os procedimentos metodológicos e os experimentos desenvolvidos ao longo deste trabalho. A investigação tem como foco a utilização direta das substituições de aminoácidos, que constituem representações explícitas das mutações observadas no genoma do SARS-CoV-2, adotadas aqui como unidade principal de análise. Na sequência, a seção de metodologia descreve as etapas de coleta e preparação dos dados, bem como as estratégias de engenharia de *features* empregadas. Por fim, a seção de experimentos apresenta a aplicação prática das técnicas de análise e classificação a partir das representações geradas por modelos baseados em *Transformers*.

3.1 Proposta

A proposta deste trabalho fundamenta-se na utilização direta das substituições de aminoácidos como forma de representação das mutações observadas no genoma do SARS-CoV-2. Essas substituições correspondem a alterações pontuais que ocorrem no processo de replicação viral, vistas na seção 1.2, e refletem modificações proteicas potencialmente associadas a efeitos funcionais relevantes, como aumento de transmissibilidade, escape imune ou impacto na eficácia vacinal. Ao adotar as substituições de aminoácidos como unidade primária de análise, busca-se capturar de forma explícita e contextualizada as variações evolutivas do vírus, superando abordagens tradicionais baseadas em sequências nucleotídicas ou restritas à proteína `Spike`.

Cada sequência de substituições é organizada como uma cadeia textual composta por *tokens* no formato `Gene:Mutação`, nos quais o gene ou proteína é identificado e associado ao código que descreve a troca de um aminoácido por outro em uma posição específica. Um exemplo clássico é a substituição `Spike:D614G`, que corresponde à troca do ácido aspártico (D) pela glicina (G) na posição 614 da proteína `Spike`. Exemplos ilustrativos por variante e por continente são apresentados nas Tabelas 3.1 e 3.2, n da Seção 3.2.1. Essa estrutura textual é compatível com modelos de linguagem, que processam os *tokens* de forma análoga a palavras em uma sentença, preservando a unidade semântica de cada mutação.

No presente trabalho, adota-se o modelo `all-MiniLM-L6-v2`, pertencente à família dos *Sentence Transformers*, para a transformação dessas cadeias em vetores densos de 384 dimensões. Cada amostra genômica, composta por uma lista de substituições, é

processada pelo modelo, que aplica mecanismos de atenção para capturar relações contextuais entre as mutações, como coocorrências frequentes ou efeitos combinatórios. O resultado é um *embedding* vetorial, que sintetiza o perfil mutacional da amostra em um espaço contínuo e de alta dimensionalidade.

A partir desses *embeddings*, propõe-se duas etapas complementares de análise. A primeira é a análise não supervisionada, realizada por meio de algoritmos de clusterização como `K-means` e `Fuzzy C-Means`, com o objetivo de explorar a estrutura latente dos dados, identificar padrões de transição evolutiva e caracterizar assinaturas mutacionais específicas de determinados grupos. A segunda etapa é a análise supervisionada, na qual modelos clássicos de aprendizado de máquina, incluindo `SVM`, `Random Forest`, `XGBoost` e `k-NN`, são aplicados para classificar automaticamente as VOCs a partir dos vetores gerados.

A inovação central da proposta reside, portanto, em considerar as substituições de aminoácidos como a unidade de entrada para modelos baseados em *Transformers*, ao invés de empregar sequências nucleotídicas completas ou focar apenas em genes específicos. Essa escolha permite preservar o significado biológico de cada mutação, ao mesmo tempo em que explora a capacidade dos modelos de atenção para representar relações contextuais entre alterações genômicas. Assim, a abordagem integra representações semânticas, clusterização e classificação supervisionada em um mesmo espaço vetorial, oferecendo uma estratégia escalável e eficaz para apoiar a vigilância genômica e o monitoramento evolutivo do SARS-CoV-2.

3.2 Metodologia

As etapas iniciais da metodologia, que são detalhadas por meio da Figura 3.1, envolveram a obtenção de dados genômicos e a curadoria das substituições de aminoácidos. As sequências foram coletadas da base GISAID em 28 de março de 2024 (Khare et al., 2021), totalizando aproximadamente 16,6 milhões de amostras. Cada entrada continha diversos metadados, entre eles a coluna `AASubstitutions`, que representa a informação central explorada neste estudo.

Com o objetivo de garantir a qualidade das análises, aplicou-se um processo de filtragem que manteve apenas as sequências marcadas como completas e com alta cobertura. A designação completa refere-se aos genomas cuja extensão cobre integralmente a sequência de referência, enquanto a alta cobertura indica que cada base foi sequenciada diversas vezes, o que reduz o risco de erro e aumenta a confiança na identificação das mutações. Além disso, foram mantidas somente as amostras classificadas como VOC, conforme as categorias estabelecidas pela OMS, que também reconhece as variantes sob monitoramento (VUM) e as variantes de interesse (VOI). As VOCs se destacam por apresentarem evidências de maior transmissibilidade, impacto na gravidade clínica ou influência na eficácia de diagnósticos, vacinas e tratamentos, sendo portanto as mais relevantes do ponto de vista da saúde pública.

Após essa etapa, o número de amostras foi reduzido para 936.638 amostras. Em seguida, foram identificadas e removidas amostras com valores extremos na quantidade de substituições de aminoácidos. Essa exclusão considerou valores abaixo do percentil

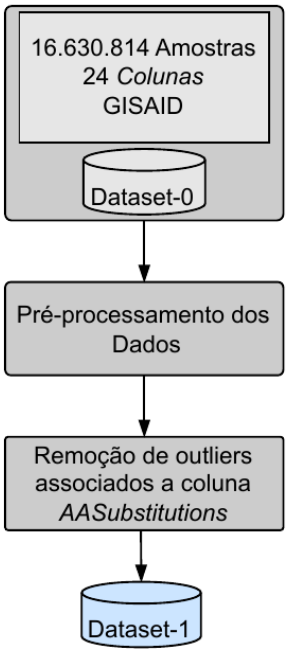


Figura 3.1: Etapas iniciais da metodologia proposta, desde a coleta dos dados genômicos no repositório GISAID até a remoção de *outliers* associados às substituições de aminoácidos, resultando na formação do *Dataset-1*

.

0,5% e acima do percentil 99,5%, buscando eliminar amostras que poderiam distorcer a distribuição vetorial. O conjunto resultante, denominado `dataset-1`, passou a conter 927.064 amostras. A Figura 3.2 ilustra a distribuição do número de substituições por amostra antes e depois da remoção de *outliers*. Inicialmente, o número de substituições variava entre 1 e 3.093, com desvio padrão próximo de 21. Após a filtragem, esse intervalo foi reduzido para valores entre 22 e 71, com desvio padrão de aproximadamente 8,8. A média foi de 35 substituições por amostra, enquanto a mediana ficou em torno de 33. Essa redução na variabilidade contribuiu para maior uniformidade e confiabilidade nas etapas subsequentes de análise.

3.2.1 Representação AASubstitutions

A coluna `AASubstitutions` apresenta, para cada amostra, a lista padronizada de substituições de aminoácidos identificadas em comparação com a sequência de referência do SARS-CoV-2. Como já explicado na subseção 2.1.1, as amostras frequentemente contêm múltiplas substituições, separadas por espaços e/ou vírgulas, formando um conjunto de eventos mutacionais associados a uma única sequência viral. Essa estrutura permite que cada amostra seja tratada como um documento composto por "palavras", sendo cada

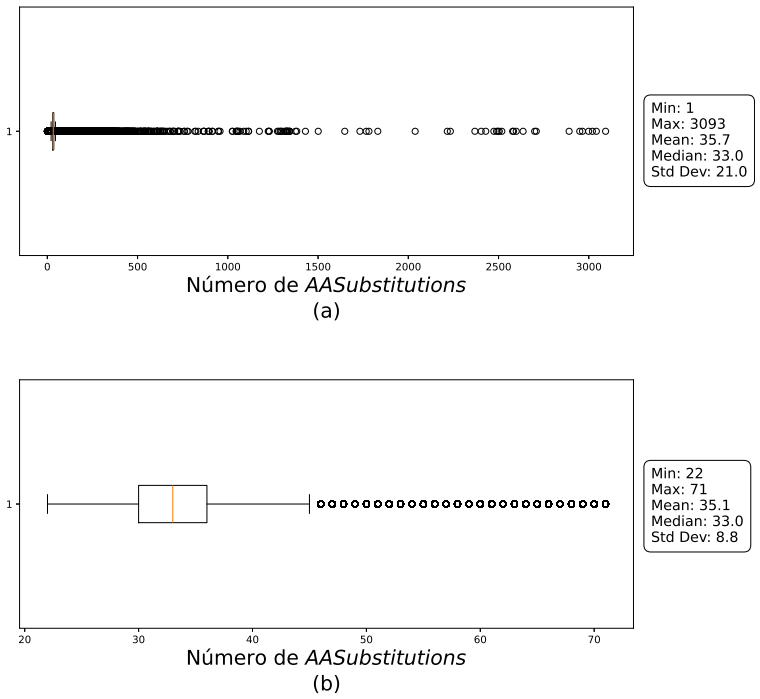


Figura 3.2: Distribuição do número de *AASubstitutions* por amostra: (a) antes da remoção de *outliers*, evidenciando alta dispersão, e (b) após a filtragem, com redução da variabilidade e maior homogeneidade do conjunto de dados.

palavra uma substituição distinta. Essa analogia viabiliza a aplicação de técnicas de PLN, como *embeddings* e modelos baseados em *Transformers*, com o objetivo de representar as sequências por meio de vetores densos que capturam padrões contextuais e semânticos entre as substituições.

Ainda que os dados apresentados nas tabelas 3.1 e 3.2 sejam padronizados, a identificação de padrões mutacionais distintos por inspeção direta é dificultada pela grande quantidade de substituições, pela variabilidade na ordem dos eventos e pela sobreposição entre variantes. O uso de modelos de linguagem baseados em *Transformers* contribui para superar essas limitações, permitindo que relações contextuais não lineares entre substituições sejam capturadas por meio de camadas de atenção.

3.2.2 Geração de Embeddings

Cada amostra genômica foi interpretada como uma sequência de substituições, representada formalmente como $s_n = [s_{n,1}, s_{n,2}, \ldots, s_{n,N_n}]$, em que $s_{n,i}$ é a i-ésima substituição da amostra n e N_n é o número total de substituições identificadas. Essa sequência textual foi então processada pelo modelo `all-MiniLM-L6-v2`, que gerou um vetor denso de 384 dimensões representado por

$$\mathbf{e}_n = [e_{n,1}, e_{n,2}, \ldots, e_{n,384}] \tag{3.1}$$

Tabela 3.1: Exemplo de amostras da variante Omicron extraídas do campo *AASubstitutions*.

Continente	*AASubstitutions*
Europa	(NSP5_P132H,Spike_H69del,Spike_T95I,Spike_A67V,Spike_S373P,Spike_Q493R, Spike_N969K,Spike_H655Y,Spike_N856K,N_R203K,Spike_G142D,NSP3_A1892T, Spike_Q954H,N_P13L,NSP3_L1266I,Spike_N501Y,N_R32del,M_Q19E,Spike_N440K, NSP4_T492I,Spike_N679K,Spike_N764K,Spike_L212I,Spike_Y505H,Spike_R346K, NSP6_I189V,Spike_T547K,M_D3G,Spike_D796Y,N_G204R,Spike_T478K, Spike_V143del,M_A63T,Spike_G496S,Spike_K417N,Spike_S371L,Spike_G339D, NSP3_S1265del,NSP14_I42V,Spike_P681H,Spike_Y144del,Spike_ins214EPE,N_S33del, Spike_S375F,Spike_Q498R,Spike_G446S,Spike_S477N,N_E31del,NSP3_K38R, Spike_N211del,Spike_E484A,E_T9I,Spike_V70del,Spike_L981F,NSP12_P323L, Spike_D614G,Spike_Y145del)
Africa	(NSP5_P132H,Spike_H69del,Spike_T95I,NSP6_A88V,Spike_A67V,NSP3_G489S, NSP4_T327I,NSP6_S106del,Spike_Q493R,Spike_N969K,Spike_H655Y,N_R203K, Spike_G142D,Spike_Q954H,N_P13L,Spike_N501Y,N_R32del,NS3_T223I,NSP14_I42V, Spike_P681H,Spike_Y144del,M_Q19E,N_S33del,NSP1_S135R,NSP4_T492I, Spike_Q498R,Spike_N679K,Spike_N764K,Spike_S477N,NSP6_G107del,Spike_Y505H, N_E31del,NSP6_F108del,Spike_D796Y,Spike_E484A,N_G204R,Spike_T478K,E_T9I, N_S413R,Spike_V70del,Spike_V143del,M_A63T,NSP12_P323L,Spike_Y145del, Spike_D614G)
Asia	(NSP5_P132H,NSP3_G489S,NSP4_T327I,Spike_S373P,Spike_Q493R,Spike_N969K, Spike_H655Y,N_R203K,Spike_V213G,Spike_G142D,Spike_Q954H,N_P13L, Spike_N501Y,NS3_T223I,Spike_T19I,M_Q19E,Spike_N440K,NSP4_T492I, Spike_N679K, Spike_N764K,Spike_Y505H,Spike_D796Y,N_G204R,Spike_T478K, N_S413R,M_A63T,Spike_S371F,NSP13_R392C,Spike_K417N,Spike_T376A, Spike_G339D,Spike_R408S,NSP4_L438F,NSP14_I42V,Spike_P681H,NSP4_L264F, NSP3_T24I,NSP1_S135R,Spike_S375F,Spike_D405N,Spike_Q498R,Spike_S477N, NSP15_T112I,Spike_E484A,E_T9I,NSP12_P323L,Spike_D614G)
America do Norte	(NSP5_P132H,Spike_H69del,Spike_T95I,Spike_A67V,NS3_S165F,Spike_S373P, Spike_Q493R,Spike_N969K,Spike_H655Y,Spike_N856K,N_R203K,Spike_G142D, NSP3_A1892T,Spike_Q954H,NS3_L106F,N_P13L,Spike_N501Y,N_R32del,M_Q19E, Spike_N440K,NSP4_T492I,NSP6_L105del,Spike_N679K,Spike_N764K,Spike_L212I, NSP6_G107del,Spike_Y505H,NSP6_I189V,Spike_T547K,M_D3G,Spike_D796Y, N_G204R,N_D343G,Spike_T478K,Spike_V143del,M_A63T,Spike_G496S,Spike_K417N, NSP6_S106del,Spike_S371L,Spike_G339D,NSP3_S1265del,NSP14_I42V,Spike_P681H, Spike_Y144del,Spike_ins214EPE,N_S33del,Spike_S375F,NSP2_A385V,Spike_Q498R, Spike_G446S,Spike_S477N,N_E31del,NSP3_K38R,Spike_N211del,Spike_E484A,E_T9I, Spike_V70del,Spike_L981F,NSP12_P323L,Spike_D614G,Spike_Y145del)
America do Sul	(NSP5_P132H,Spike_T95I,Spike_S373P,Spike_Q493R,Spike_N969K,Spike_H655Y, Spike_N856K,N_R203K,Spike_G142D,NSP3_A1892T,NSP3_P1103S,Spike_Q954H, N_P13L,NSP3_L1266I,Spike_N501Y,M_Q19E,Spike_N440K,NSP4_T492I,Spike_N679K, Spike_N764K,Spike_Y505H,NSP6_I189V,Spike_T547K,M_D3G,Spike_D796Y,N_G204R, Spike_T478K,Spike_V143del,M_A63T,Spike_G496S,Spike_ins215PEX,Spike_K417N, Spike_S371L,Spike_G339D,NSP3_S1265del,NSP14_I42V,Spike_P681H,Spike_Y144del, Spike_S375F,Spike_Q498R,Spike_G446S,Spike_D215E,Spike_S477N,NSP3_K38R, Spike_E484A,E_T9I,Spike_L981F,NSP12_P323L,Spike_D614G,Spike_Y145del)

Tabela 3.2: Exemplo de amostras da variante Delta extraídas do campo *AASubstitutions*.

Continente	*AASubstitutions*
Europa	(N_G215C,N_D63G,N_R203M,NSP12_G671S,Spike_G142D,NS3_S26L,NSP3_R1341Q, Spike_P681R,Spike_R158del,NSP3_P1228L,NS6_K42N,NS7a_V82A,NSP3_A488S, Spike_F157del,NSP4_T492I,NSP14_A394V,Spike_T19R,NS7a_T120I,N_D377Y,M_I82T, Spike_D950N,NS7b_T40I,NSP13_P77L,Spike_E156G,NSP3_P1469S,Spike_T478K, NSP6_T77A,NSP12_K807E,NSP6_A2V,NSP12_P323L,Spike_D614G,NSP4_V167L, Spike_L452R)
Africa	(N_G215C,N_D63G,N_R203M,NSP12_G671S,Spike_G142D,NS3_S26L,Spike_P681R, Spike_R158del,NSP3_P1228L,NS7a_V82A,NSP2_D511N,NSP3_A488S,Spike_F157del, NSP4_T492I,NSP14_A394V,Spike_T19R,NS7a_T120I,N_D377Y,M_I82T,Spike_D950N, NS7b_T40I,NSP13_P77L,Spike_E156G,NSP3_P1469S,Spike_T478K,NSP6_T77A, NS8_G66S,NSP12_P323L,Spike_D614G,NSP4_V167L,Spike_L452R,NSP13_E365D)
Asia	(N_G215C,Spike_T95I,N_D63G,N_R203M,NSP12_G671S,NSP15_Q19H,NS3_S26L, NS3_T221K,Spike_P681R,Spike_R158del,NSP3_P1228L,Spike_D950B,NS7a_V82A, NSP3_A488S,Spike_F157del,NSP4_T492I,Spike_T19R,NSP14_A394V,NS7a_T120I, N_D377Y,M_I82T,NSP3_A231V,NS7b_T40I,NSP13_P77L,Spike_E156G,NSP14_N67S, NSP3_P1469S,Spike_T478K,NSP6_T77A,NSP12_P323L,NS3_G49V,Spike_D614G, NSP4_V167L,Spike_L452R,Spike_P26L)
America do Norte	(N_G215C,N_D63G,N_R203M,NSP12_G671S,NSP13_I334V,Spike_G142D,NS3_S26L, Spike_P681R,Spike_R158del,NSP3_P1159L,NSP3_P1228L,NS7a_V82A,NSP3_A488S, Spike_F157del,NS8_P36S,NSP4_T492I,NSP14_A394V,Spike_T19R,NS7a_T120I, N_D377Y,M_I82T,Spike_D950N,NS7b_T40I, NSP13_P77L,Spike_E156G,NSP3_P1469S, NSP3_S1877N,Spike_T478K,NSP6_T77A,NSP12_P323L,Spike_D614G,NSP4_V167L, Spike_L452R)
America do Sul	(N_G215C,N_D63G,N_R203M,NSP8_T145I,NSP12_G671S,NS3_S26L,Spike_P681R, Spike_R158del,NSP3_P1228L,NS7a_V82A,NSP3_A488S,Spike_F157del,NSP4_T492I, NSP14_A394V,Spike_T19R,NS7a_T120I,N_D377Y,M_I82T,Spike_D950N,NS7b_T40I, NSP13_P77L,Spike_E156G,NSP3_P1469S,Spike_T478K,NSP6_T77A,NSP12_P323L, Spike_D614G,Spike_L452R,NSP4_V167L,NSP3_I273V)

em que $\mathbf{e}_n$ corresponde ao embedding vetorial da amostra n.

Para garantir unicidade no conteúdo informativo das amostras, foi criado um subconjunto denominado `Dataset-2`, contendo apenas as entradas únicas com base na coluna `AASubstitutions`, o que resultou em 550.088 sequências distintas. Esse conjunto foi utilizado como entrada no modelo *Transformer*.

Após a vetorização, verificou-se que muitos *embeddings* apresentavam representações quase idênticas, reflexo da elevada similaridade entre os perfis mutacionais das amostras. Esse fenômeno levou à eliminação da redundância entre os vetores de *embedding*, resultando em um conjunto final composto por 317.319 amostras únicas, tanto em termos de substituições de aminoácidos quanto de representação vetorial.

Essa similaridade entre amostras pode ser exemplificada na Tabela 3.3, que apresenta duas sequências distintas oriundas do conjunto original, mas com alta sobreposição de mutações, a primeira difere da segunda por apenas três substituições, sendo uma exclusiva da primeira amostra e duas exclusivas da segunda. Essa pequena divergência ao longo da

sequência é suficiente para manter os vetores gerados praticamente indistinguíveis em um espaço de alta dimensionalidade, o que justifica a necessidade da filtragem de *embeddings* redundantes. Todo *pipeline* do processo descrito é ilustrado na Figura 3.3.

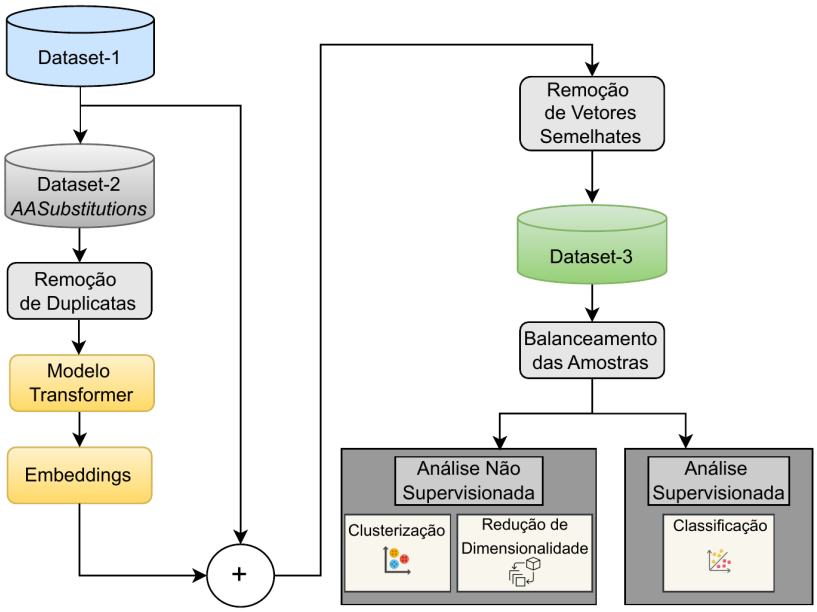


Figura 3.3: Etapas finais da metodologia, ilustrando a integração dos *embeddings* gerados ao *dataset-1* e a construção do *dataset-3*, utilizado nas análises vetoriais e nos experimentos subsequentes.

A fim de mitigar possíveis vieses causados pelo desbalanceamento entre as VOCs, aplicou-se uma estratégia de subamostragem, no `Dataset-3`, com base na menor classe, correspondente à variante Beta, composta por 5.865 amostras. Como resultado, obteve-se um conjunto balanceado contendo 29.325 amostras distribuídas igualmente entre as variantes Alpha, Beta, Gamma, Delta e Omicron. A matriz de *embeddings* produzida ao final desse processo serviu como base para as análises de clusterização, conduzidas por meio dos algoritmos `K-means` e `Fuzzy C-Means`, e também para os experimentos de classificação supervisionada.

3.3 Experimentos

Esta seção apresenta os experimentos realizados com os *embeddings* gerados a partir das substituições de aminoácidos, após o pré-processamento das amostras genômicas. As análises envolveram técnicas de agrupamento, redução de dimensionalidade e classificação supervisionada, com o objetivo de investigar a organização estrutural dos dados no espaço vetorial e identificar padrões associados às variantes do SARS-CoV-2.

As abordagens de clusterização foram utilizadas para detectar agrupamentos coerentes e possíveis zonas de transição entre variantes, enquanto os métodos de projeção serviram

para representar visualmente a estrutura latente dos dados. Os modelos supervisionados, por sua vez, permitiram avaliar a capacidade dos *embeddings* em discriminar automaticamente entre diferentes variantes virais com base em seus perfis mutacionais. Em conjunto, os experimentos buscaram validar a utilidade das representações vetoriais na caracterização genômica das amostras.

Os experimentos foram executados em uma estação de trabalho com processador Intel Core i9-13900K, 128 GB de memória RAM e sistema operacional Windows 11 Pro, 64 bits. Para aceleração computacional, foi utilizada uma GPU NVIDIA GeForce RTX 3060, compatível com CUDA, viabilizando a execução de modelos baseados em Transformers.

O ambiente de desenvolvimento foi construído em Python 3.11, com ambientes virtuais configurados via `venv`. As bibliotecas utilizadas incluíram `NumPy` e `Pandas` para manipulação de dados, `Scikit-Learn` e `CuML` para tarefas de aprendizado de máquina, `PyTorch` para deep learning, `Transformers` da *Hugging Face* para uso de modelos pré-treinados, além de `Matplotlib` e `Seaborn` para visualizações.

3.3.1 Agrupamento com o K-means

O algoritmo `K-means` foi aplicado diretamente sobre os *embeddings* obtidos a partir das substituições de aminoácidos. O objetivo foi identificar agrupamentos latentes no espaço vetorial das amostras virais.

Uma característica importante e obrigatória na utilização desse algoritmo é a determinação do número ideal de *clusters*. Para esse fim, foram utilizadas quatro métricas de avaliação: `método do cotovelo` (*Elbow*), `Silhouette Score`, `índice de Calinski-Harabasz` e `índice de Davies-Bouldin`. A figura 3.4 mostra o gráfico resultante do método do cotovelo, enquanto a tabela 3.4 apresenta os valores obtidos como resultado da aplicação dos demais métodos.

Analisando os resultados obtidos, observamos que o `método do cotovelo`, o índice de `Silhouette` e o índice de `Davies-Bouldin` apontaram $k = 5$ como a configuração mais adequada, indicando o melhor equilíbrio entre compactação e separação dos grupos. Embora o índice de `Calinski-Harabasz` não tenha identificado $k = 5$ como valor ótimo, esse valor apareceu como o segundo melhor resultado segundo essa métrica. Diante desse conjunto de evidências, adotou-se a configuração com cinco *clusters* para os experimentos subsequentes, voltados à visualização e à análise estrutural das variantes virais.

3.3.2 Agrupamento com Fuzzy C-means

Embora o `K-means` tenha se mostrado eficaz na identificação de agrupamentos coerentes com as variantes virais, sua principal limitação está na atribuição rígida de cada amostra a um único *cluster*. Essa abordagem desconsidera possíveis sobreposições mutacionais entre linhagens, bem como transições graduais que frequentemente ocorrem no contexto evolutivo do SARS-CoV-2. Dada a natureza contínua e altamente combinatória das substituições de aminoácidos, espera-se que algumas amostras apresentem perfis híbridos ou ambíguos, que não se ajustam bem a uma segmentação binária do espaço vetorial. Nesse sentido, a escolha pelo algoritmo `FCM` representa uma evolução metodo-

Tabela 3.3: Comparação entre duas amostras semelhantes, com as substituições de aminoácidos distintas realçadas em vermelho.

AASubstitutions
(NSP5_P132H, Spike_H69del, Spike_T95I, Spike_A67V, Spike_S373P, Spike_Q493R, Spike_N969K, Spike_H655Y, Spike_N856K, N_R203K, Spike_G142D, NSP3_A1892T, Spike_Q954H, N_P13L, NSP3_L1266I, Spike_N501Y, N_R32del, M_Q19E, Spike_N440K, NSP4_T492I, NSP6_L105del, Spike_N679K, Spike_N764K, NSP6_G107del, Spike_R346K, NSP6_I189V, Spike_T547K, M_D3G, Spike_D796Y, N_G204R, Spike_T478K, Spike_V143del, M_A63T, Spike_G496S, Spike_K417N, NSP6_S106del, Spike_S371L, Spike_G339D, NSP3_S1265del, NSP14_I42V, Spike_P681H, Spike_Y144del, N_S33del, Spike_S375F, Spike_Q498R, Spike_G446S, Spike_S477N, N_E31del, NSP3_K38R, Spike_E484A, E_T9I, Spike_V70del, Spike_L981F, NSP12_P323L, Spike_D614G, Spike_Y145del)
(NSP5_P132H, Spike_H69del, Spike_T95I, Spike_A67V, Spike_S373P, Spike_Q493R, Spike_N969K, Spike_H655Y, Spike_N856K, N_R203K, Spike_G142D, NSP3_A1892T, Spike_Q954H, N_P13L, NSP3_L1266I, Spike_N501Y, N_R32del, N_T49I, M_Q19E, Spike_N440K, NSP4_T492I, NSP6_L105del, Spike_N679K, Spike_N764K, NSP6_G107del, Spike_Y505H, NSP6_I189V, Spike_T547K, M_D3G, Spike_D796Y, N_G204R, Spike_T478K, Spike_V143del, M_A63T, Spike_G496S, Spike_K417N, NSP6_S106del, Spike_S371L, Spike_G339D, NSP3_S1265del, NSP14_I42V, Spike_P681H, Spike_Y144del, N_S33del, Spike_S375F, Spike_Q498R, Spike_G446S, Spike_S477N, N_E31del, NSP3_K38R, Spike_E484A, E_T9I, Spike_V70del, Spike_L981F, NSP12_P323L, Spike_D614G, Spike_Y145del)

Tabela 3.4: Avaliação do número de *clusters* com base em métricas internas de qualidade do `k-means`, sendo os valores em vermelho aqueles que representam o melhor desempenho em cada métrica.

Número de *Clusters*	**Índice de *Silhouette***	**Índice de *Calinski-Harabasz***	**Índice de *Davies-Bouldin***
2	0,33	11.970,76	1,15
3	0,27	11.046,63	1,57
4	0,33	11.507,18	1,20
5	0,36	11.790,32	1,14
6	0,34	10.274,33	1,59
7	0,32	9.082,49	1,61
8	0,29	8.263,20	1,68
9	0,29	7.656,35	1,72
10	0,26	7.108,78	1,87

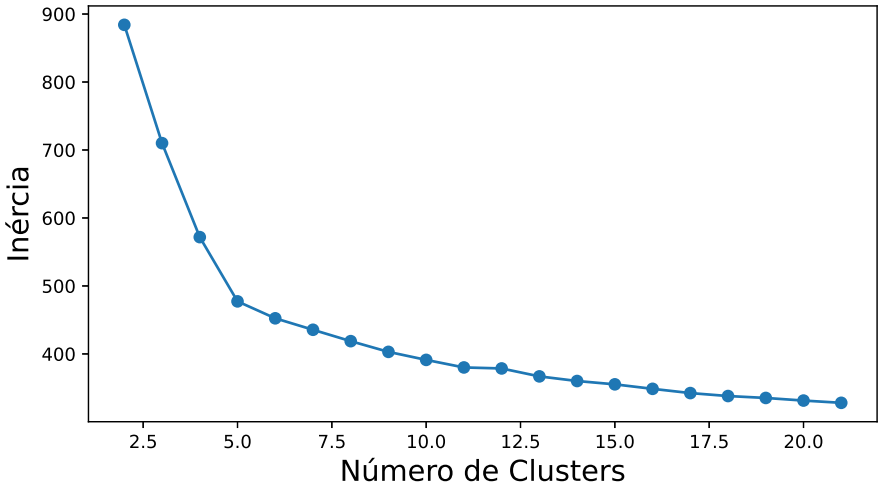


Figura 3.4: Aplicação do método do cotovelo (*Elbow*) para avaliação do número adequado de clusters, com base na variação da soma dos erros quadráticos intra-cluster.

lógica, ao permitir que cada amostra seja representada como uma combinação ponderada de múltiplos *clusters*, com graus de associação contínuos.

Do ponto de vista computacional, o `FCM` oferece uma modelagem mais flexível e refinada do espaço latente gerado pelos *embeddings*. Em vez de impor fronteiras rígidas entre grupos, o `FCM` calcula a pertinência de cada amostra a cada *cluster* com base na distância relativa aos centróides, utilizando um expoente de fuzzificação que controla a suavidade da segmentação. Essa propriedade é valiosa para capturar zonas de ambiguidade evolutiva. Além disso, a matriz de pertinência *fuzzy* pode ser explorada analiticamente para identificar amostras de baixa confiança, mutações compartilhadas entre *clusters* ou padrões de exclusividade funcional, ampliando o potencial interpretativo do modelo além da classificação rígida oferecida pelo `K-means`.

O experimento avaliou valores de K, assim como no método `k-means`. Para cada configuração, foram calculadas métricas *fuzzy* clássicas, incluindo a `PE`, o `FPC` e o índice `XB`, com o intuito de selecionar o número de *clusters* que proporcionasse o melhor equilíbrio entre compactação *intra-cluster* e separabilidade entre grupos. A Tabela 3.5 apresenta os resultados obtidos para cada valor de K nas três métricas *fuzzy* analisadas. O valor $K = 5$ obteve o melhor desempenho no índice de `XB`, e o segundo melhor resultado na `PE`. Embora não tenha se destacado de forma isolada em todas as métricas, seu desempenho equilibrado entre os critérios avaliados, somado à consistência com a configuração adotada nos experimentos com o algoritmo `K-means`, motivou a escolha de $K = 5$ como valor final para os experimentos posteriores com o `FCM`.

A execução do `FCM` resultou em uma matriz de pertinência de dimensão $N \times K$, na qual cada elemento representa o grau de associação de uma amostra ao respectivo *cluster*. A partir dessa matriz, foi derivada uma variável auxiliar denominada `ClusterMax`, responsável por indicar, para cada amostra, o *cluster* com maior grau de pertinência. Essa variável permitiu a comparação com os agrupamentos obtidos pelo `K-means`, além de possibilitar visualizações em espaço reduzido com cores atribuídas conforme a maior

Tabela 3.5: Avaliação do número de *clusters* com base em métricas internas de qualidade do `Fuzzy C-means`, sendo os valores em vermelho aqueles que representam o melhor desempenho em cada métrica.

Número de *Clusters*	**FCP**	**PE**	**XB**
2	0,55	0,63	1,02
3	0,43	1,22	1,93
4	0,39	1,21	0,65
5	0,40	1,13	0,35
6	0,32	1,60	0,61
7	0,27	1,75	$1{,}57e+10$
8	0,24	1,89	$1{,}78e+12$
9	0,21	2,0	$1{,}045e+13$
10	0,18	2,2	$3{,}67e+11$

associação *fuzzy*.

A matriz de pertinência foi utilizada também para classificar as amostras segundo a confiabilidade de sua associação. Aqueles casos com baixos valores máximos de pertinência foram considerados ambíguos, e analisados quanto ao seu perfil de distribuição *fuzzy* e às substituições de aminoácidos predominantes. Além disso, foi conduzida uma análise de exclusividade por mutação, considerando a agregação dos graus de pertinência de todas as amostras em que cada substituição aparece. Esse procedimento permitiu identificar mutações mais específicas de determinados *clusters* e outras compartilhadas entre múltiplos grupos, oferecendo uma visão mais refinada da estrutura genômica das amostras.

3.3.3 Redução de Dimensionalidade com t-SNE

Para possibilitar a visualização dos dados de alta dimensionalidade, foi empregada a técnica de redução de dimensionalidade `t-SNE`. Essa técnica não linear permite representar os *embeddings* em um espaço bidimensional, com o objetivo de preservar as relações locais de vizinhança entre os dados.

A aplicação do algoritmo `t-SNE` requer a definição de alguns parâmetros, conforme discutido na seção 2.2.1 e detalhado na tabela 3.6. Esse parâmetros influenciam diretamente a qualidade da projeção e a fidelidade estrutural da visualização. O número de componentes foi fixado em 2, possibilitando a representação gráfica bidimensional das amostras no plano cartesiano. A perplexidade foi ajustada para o valor 30,0, escolha fundamentada na recomendação van der Maaten e Hinton (2008), que sugerem valores típicos entre 5 e 50, dependendo do tamanho do conjunto de dados. O valor 30 é amplamente utilizado na literatura como ponto de equilíbrio entre a preservação das estruturas locais e globais, tendo apresentado bons resultados preliminares nos testes exploratórios realizados neste trabalho.

A taxa de aprendizado controla a magnitude dos ajustes realizados nas posições dos

pontos ao longo da otimização. Esse valor visa garantir estabilidade no processo, evitando tanto a convergência prematura quanto flutuações excessivas. A métrica de distância adotada foi a `Cosseno`, escolhida por sua adequação a dados de alta dimensionalidade, como os *embeddings* utilizados neste estudo. Essa métrica avalia a similaridade angular entre vetores, tornando-se menos sensível à magnitude absoluta e mais focada no padrão relativo de variação entre dimensões. Tal característica é relevante em *embeddings*, nos quais a informação semântica é representada predominantemente pela direção no espaço vetorial, sendo a distância `Cosseno` amplamente empregada em aplicações de PLN e recuperação de informação para medir similaridade entre representações vetoriais.

O método de inicialização da projeção foi baseado em `PCA`, uma técnica que provê um ponto de partida estruturado para o algoritmo, contribuindo para a estabilidade e reprodutibilidade da visualização. Por fim, a semente aleatória foi definida para garantir a reprodutibilidade dos resultados, aspecto essencial para a consistência metodológica em pesquisas científicas.

Tabela 3.6: Parâmetros utilizados no algoritmo t-SNE.

Parâmetro	**Valor**
`Número de componentes`	2
`Perplexidade`	`30.0`
`Taxa de aprendizado`	200.00
`Amplificação das distâncias`	`12.0`
`Métrica de distância`	'cosine'
`Método de inicialização`	'pca'
`Semente aleatória`	42

Além da projeção bidimensional obtida com o `t-SNE`, foram gerados histogramas individuais para cada uma das dimensões resultantes, conforme ilustrado na Figura 3.5. O objetivo dessa visualização é analisar a distribuição das amostras ao longo de cada eixo da projeção, permitindo a identificação de regiões com maior concentração de dados.

A presença de múltiplos picos nas distribuições das duas dimensões projetadas pelo `t-SNE` indica a existência de agrupamentos locais ou subestruturas internas no espaço vetorial original. Como o `t-SNE` preserva as relações de proximidade entre vetores semelhantes durante a redução de dimensionalidade, a concentração de amostras em regiões específicas do espaço bidimensional sugere que certos subconjuntos compartilham padrões mutacionais similares. Essa característica é relevante para as análises de clusterização subsequentes, pois reforça a hipótese de que há estrutura latente nos dados.

Além disso, a distribuição das amostras ao longo dos eixos da projeção evidencia uma dispersão adequada no plano gerado. Não se observa compressão excessiva nem dispersão aleatória, o que indica que os parâmetros adotados para o `t-SNE` resultaram em uma projeção estável. Esse comportamento sugere que a técnica foi eficaz em preservar a estrutura local dos dados, favorecendo a visualização e a interpretação das relações de similaridade entre amostras no espaço reduzido.

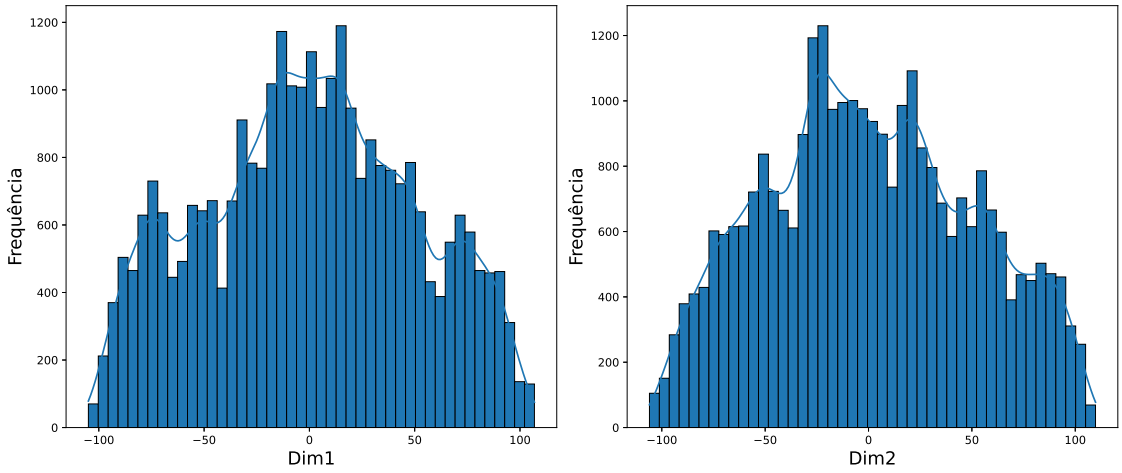


Figura 3.5: Distribuição das amostras nas duas dimensões resultantes da projeção via t-SNE. A Figura (a) apresenta a distribuição ao longo da primeira dimensão (Dim1), enquanto a Figura (b) corresponde à segunda dimensão (Dim2), evidenciando a estrutura dos dados no espaço reduzido.

3.3.4 Analise Supervisionada: Classificação

A etapa de classificação supervisionada teve como objetivo avaliar o potencial discriminativo dos embeddings gerados a partir das substituições de aminoácidos na tarefa de identificação das variantes virais. Para isso, foram selecionados quatro algoritmos clássicos de aprendizado supervisionado: `SVM` com *kernel* radial, `Random Forest`, `XGBoost` e `k-NN`. Todos os modelos foram treinados com os *embeddings* previamente balanceados entre as variantes Alpha, Beta, Gamma, Delta e Omicron.

Durante o treinamento, foram identificadas duas dimensões invariantes nos vetores de *embedding*, correspondentes aos índices 224 e 320, que apresentavam variância nula em todas as amostras. Essas dimensões foram removidas por não contribuírem à discriminação entre as classes e por poderem introduzir ruído em modelos sensíveis à redundância, como o `SVM` e o `k-NN`. Após a limpeza, os vetores passaram a conter 382 atributos informativos.

Os experimentos supervisionados foram conduzidos utilizando validação cruzada estratificada do tipo `k-fold`, com $k = 5$, assegurando uma avaliação consistente e imparcial do desempenho preditivo. Durante esse processo, foram ajustados hiperparâmetros específicos para cada modelo. Para o `SVM`, variou-se o parâmetro de penalização C; na `Random Forest`, testaram-se diferentes valores de $mtry$; no `XGBoost`, exploraram-se combinações entre profundidade de árvore, fração de amostras por árvores e número de iterações; e no `k-NN`, diferentes valores de k foram avaliados. A seleção dos melhores parâmetros foi feita com base na minimização do *logLoss*, buscando garantir não apenas maior acurácia, mas também confiabilidade nas probabilidades estimadas.

Além das predições geradas durante a validação cruzada, foi construído um conjunto de teste externo composto por amostras que não foram utilizadas durante o balanceamento. Esse conjunto incluiu um total de 287.994 amostras, distribuídas entre as vari-

antes Alpha, Delta, Gamma e Omicron. A acurácia foi utilizada como métrica principal para avaliação nesse conjunto, sendo calculada separadamente para cada variante. Essa abordagem permitiu analisar a capacidade de generalização dos modelos frente a dados não vistos e com distribuição original desbalanceada.

Capítulo 4

Resultados e Discussões

Este capítulo, são apresentados e discutidos os resultados obtidos a partir dos experimentos descritos no Capítulo 3, integrando múltiplas abordagens para análise de perfis mutacionais do SARS-CoV-2 a partir de representações vetoriais derivadas de substituições de aminoácidos. Os experimentos foram organizados em três eixos principais: (i) análise não supervisionada com projeções `t-SNE` e agrupamentos via `K-means`; (ii) análise não supervisionada com projeções `t-SNE` e agrupamentos via `Fuzzy C-Means`; e (iii) classificação supervisionada de variantes virais com modelos baseados em árvores de decisão e redes de suporte vetorial.

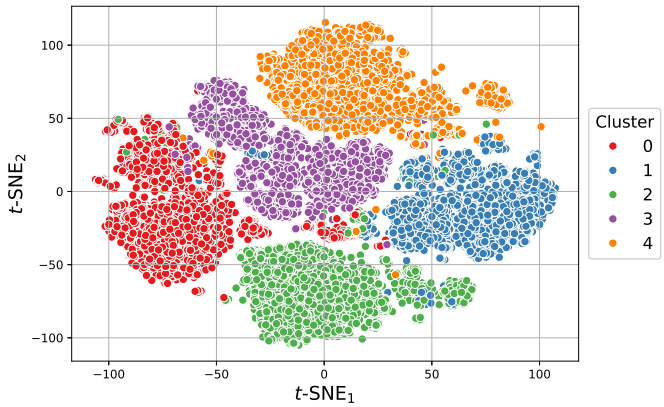
A ideia central desta seção é verificar se os agrupamentos formados a partir dos *embeddings* refletem padrões coerentes com informações epidemiológicas e genéticas conhecidas, como a distribuição VOC e assinaturas mutacionais específicas. Além disso, busca-se avaliar a eficácia da abordagem vetorial tanto do ponto de vista exploratório quanto preditivo, com foco na interpretabilidade biológica e na escalabilidade computacional.

4.1 Resultados com Base nos Agrupamentos K-Means

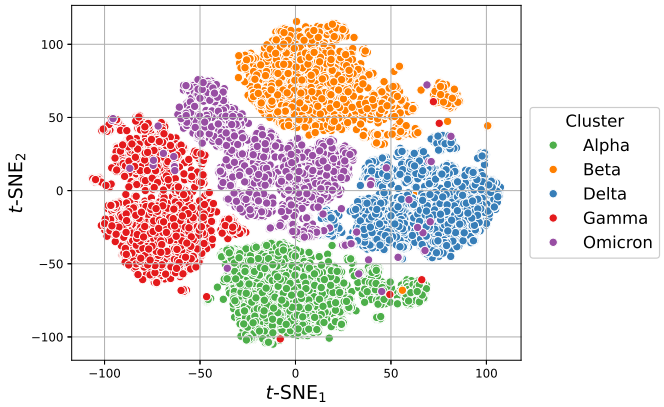
A aplicação do algoritmo `K-Means` sobre os *embeddings* de substituições de aminoácidos resultou na formação de cinco agrupamentos distintos. Para auxiliar a visualização da estrutura latente desses agrupamentos no espaço vetorial original, foi aplicada a técnica de redução de dimensionalidade `t-SNE`, projetando os dados em duas dimensões. A Figura 4.1a apresenta a projeção `t-SNE` colorida de acordo com os *clusters* atribuidos pelo `k-Means`, enquanto a Figura 4.1b mostra a mesma projeção, desta vez colorida pelas VOCs.

Visualmente, observa-se uma forte correspondência entre os agrupamentos formados e as classes reais das variantes. As regiões bem definidas indicam que o modelo foi capaz de agrupar as amostras de forma coerente com suas assinaturas mutacionais, com boa separabilidade entre os grupos, mesmo em um espaço de alta dimensionalidade.

A Tabela 4.1 detalha a contagem absoluta e percentual de cada VOC dentro de cada agrupamento. A análise desses dados revela que o *Cluster* 0 é composto majoritariamente por amostras da variante Gamma, que representa 93,45% do grupo, seguido por uma pequena proporção de Omicron (5,28%). O *Cluster* 1 é dominado pela variante Delta



(a) Projeção colorida por cluster (`K-Means`).



(b) Projeção colorida por variante (VOC).

Figura 4.1: Projeção das amostras obtido via *t*-SNE e agrupamento *k*-Means.

(97,26%), com pequenas contribuições de Alpha, Omicron, Beta e Gamma. Já o *Cluster* 2 concentra principalmente a variante Alpha, que responde por 94,94% das amostras, com Delta (2,50%) e Gamma (1,74%) aparecendo em proporções reduzidas. O *Cluster* 3 apresenta uma predominância clara da variante Omicron, que corresponde a 98,75% das amostras, enquanto o *Cluster* 4 é praticamente exclusivo da variante Beta, com 98,29%.

A coerência observada nos agrupamentos formados pelo `K-Means` também pode ser interpretada com base na biologia molecular das variantes. A predominância da variante Delta no *cluster* 1 está associada à presença de mutações características como `Spike:T19R` e `P681R`, que aumentam a afinidade com o receptor ACE2 e conferem resistência parcial à neutralização por anticorpos monoclonais (Cherian et al., 2021). Da mesma forma, a concentração da variante Alpha reflete a influência de substituições como `Spike:A570D`, `P681H` e `D1118H`, relacionadas ao aumento de infectividade e potencial evasão imune (Ladhani et al., 2021).

A variante Gamma, marcada pelas mutações `Spike:H665Y`, `L18F` e `T1027I`, tende a formar agrupamentos coesos devido à atuação combinada dessas substituições no domínio de ligação ao receptor (Faria et al., 2021). Já a Beta, apresenta um perfil mutacional

Tabela 4.1: Distribuição de variantes VOC por cluster.

Cluster	Alpha	Beta	Delta	Gamma	Omicron
0	5 (0,08%)	33 (0,55%)	39 (0,64%)	5677 (93,45%)	321 (5,28%)
1	80 (1,39%)	10 (0,17%)	5603 (97,26%)	4 (0,07%)	64 (1,11%)
2	5775 (94,94%)	26 (0,43%)	152 (2,50%)	106 (1,74%)	24 (0,39%)
3	1 (0,02%)	6 (0,10%)	13 (0,24%)	49 (0,89%)	5446 (98,75%)
4	4 (0,07%)	5790 (98,29%)	58 (0,98%)	29 (0,49%)	10 (0,17%)

altamente específico, incluindo `Spike:A701V` e `E484K`, que justifica sua concentração em regiões vetoriais restritas (Tegally et al., 2021). Por fim, a Omicron, com mais de 30 alterações na proteína `Spike`, incluindo `Q954H`, `N679K` e `D796Y`, forma *clusters* bem definidos em virtude de sua assinatura mutacional singular (Bhattacharya et al., 2022). Todas essas mutações estão de acordo com a Figura 4.2, que apresenta as 10 mutações mais frequentes em cada *cluster*, reforçando que os agrupamentos refletem padrões mutacionais recorrentes e biologicamente significativos.

Apesar dessa organização clara, observa-se alguma sobreposição entre certos *clusters*, particularmente entre Gamma e Omicron. Essa transição pode ser explicada pela presença de mutações recorrentes entre variantes, como `Spike:D614G`, `N501Y` e `E484K`, que ocorrem em múltiplas linhagens. Essa sobreposição também é evidenciada pela Figura 4.2, onde se observa que diversas mutações aparecem em mais de um *cluster*. Essa recorrência pode ser atribuída à atuação de pressões seletivas convergentes, que favorecem alterações com impacto funcional positivo, como maior transmissibilidade ou escape imune (Câmara, 2024). Tais características geram regiões vetoriais próximas no espaço de *embeddings*, o que limita a capacidade do `K-Means` de segmentar amostras em fronteiras bem definidas.

Ainda assim, a análise dos agrupamentos obtidos por `K-Means` fornece uma primeira evidência de que os *embeddings* baseados em substituições de aminoácidos contêm informações discriminativas suficientes para a separação não supervisionada de variantes virais. Os agrupamentos latentes identificados, embora limitados pela rigidez do modelo, estabelecem uma base sólida para análises supervisionadas e para a identificação de padrões evolutivos em estudos e análises posteriores.

4.2 Resultados com Base nos Agrupamentos Fuzzy C-Means

Nesta seção, são apresentados os resultados obtidos com a aplicação do algoritmo FCM sobre os *embeddings* genômicos derivados das substituições de aminoácidos. Diferentemente da clusterização rígida, o FCM representa graus de pertinência das amostras a múltiplos grupos, oferecendo uma visão mais flexível da estrutura latente dos dados. Essa abordagem complementa o `K-Means`, que apresenta limitações para lidar com zonas de transição e sobreposição entre variantes geneticamente próximas. As análises a seguir investigam a coerência biológica dos agrupamentos *fuzzy*, a presença de ambiguidade

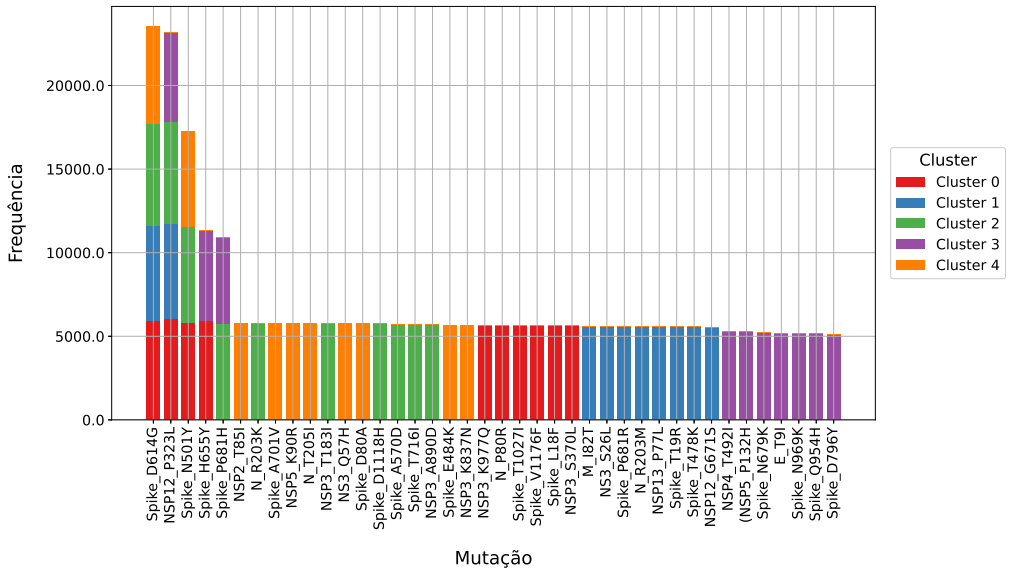


Figura 4.2: Mutações mais frequentes associadas a cada cluster obtido com o algoritmo K-Means, destacando padrões mutacionais característico dos agrupamentos identificados.

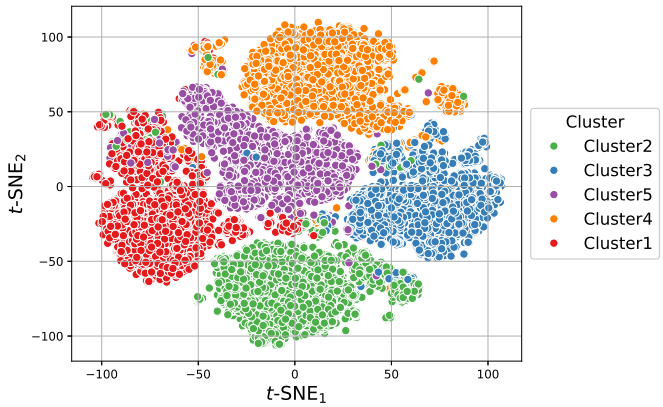
mutacional e a identificação de assinaturas específicas associadas a diferentes variantes virais.

4.2.1 Caracterização e Validação dos Agrupamentos Fuzzy

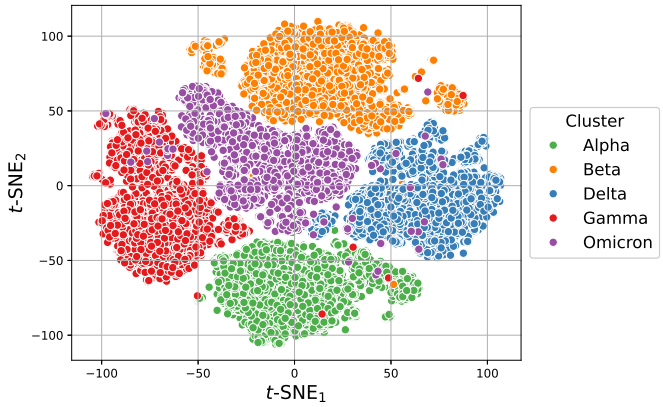
A análise inicial da estrutura dos agrupamentos, obtidos pelo algoritmo `FCM`, foi realizada por meio de uma projeção bidimensional dos *embeddings* genômicos, utilizando a técnica de redução de dimensionalidade `t-SNE`. A Figura 4.3a ilustra a distribuição das amostras no espaço latente, onde cada ponto é colorido segundo o *cluster* de maior grau de pertinência `ClusterMax`. A visualização revela a formação de regiões bem definidas, com diferentes níveis de separabilidade, o que constitui uma evidência primária de que os *embeddings* capturam eficazmente padrões mutacionais com potencial discriminativo.

Para avaliar a correspondência entre a estrutura de agrupamento induzida pelo modelo e a classificação epidemiológica consolidada, a mesma projeção `t-SNE` foi colorida de acordo com as VOCs, conforme exibido na Figura 4.3b. Embora se observe uma sobreposição natural entre as variantes, esperada devido à ancestralidade compartilhada, nota-se que linhagens específicas tendem a ocupar regiões distintas no espaço de projeção. Esse alinhamento sugere uma consistência notável entre o agrupamento não supervisionado e a separação biológica conhecida das variantes.

A fim de quantificar essas associações, foi gerado um mapa de calor, ilustrado pela Figura 4.4, que representa a média de pertinência *fuzzy* de cada VOC para cada *cluster*. Esta análise quantitativa corrobora com as observações da projeção `t-SNE`. Identificou-se que determinadas variantes demonstram uma forte associação com um único agrupamento, como a Alpha, com 51,3% de sua pertinência média concentrada no *Cluster* 1; a Delta, com 68,7% no *Cluster* 2; e a Beta, com 55,5% no *Cluster* 3. Este forte alinhamento indica



(a) Projeção colorida por cluster (`ClusterMax`).



(b) Projeção colorida por variante (VOC).

Figura 4.3: Projeção das amostras no espaço latente bidimensional obtido via *t*-SNE.

que o modelo conseguiu isolar perfis mutacionais característicos dessas VOCs.

Em contrapartida, outras variantes como a Gamma e a Omicron apresentam uma distribuição de pertinência mais difusa, com associações relevantes em múltiplos *clusters*. A Gamma, por exemplo, distribui-se de forma predominante entre o *Cluster* 4 (53,6%) e o *Cluster* 1 (15,4%), enquanto a Omicron, embora dominante no *Cluster* 5 (51,6%), também exibe conexões com outros grupos. Tal comportamento sugere a presença de padrões mutacionais mais heterogêneos ou transicionais. Esses resultados evidenciam a principal vantagem da abordagem *fuzzy*, que é sua capacidade de modelar tanto assinaturas genômicas específicas quanto zonas de sobreposição entre variantes evolutivamente próximas, sem impor fronteiras rígidas e artificiais.

4.2.2 Identificação de Amostras Ambíguas e Perfis Transicionais

Uma das contribuições centrais da clusterização *fuzzy* reside na sua habilidade de quantificar a incerteza na classificação de cada amostra. Para explorar essa característica, calculou-se o valor máximo de pertinência `PertinenciaMax` de cada amostra em relação

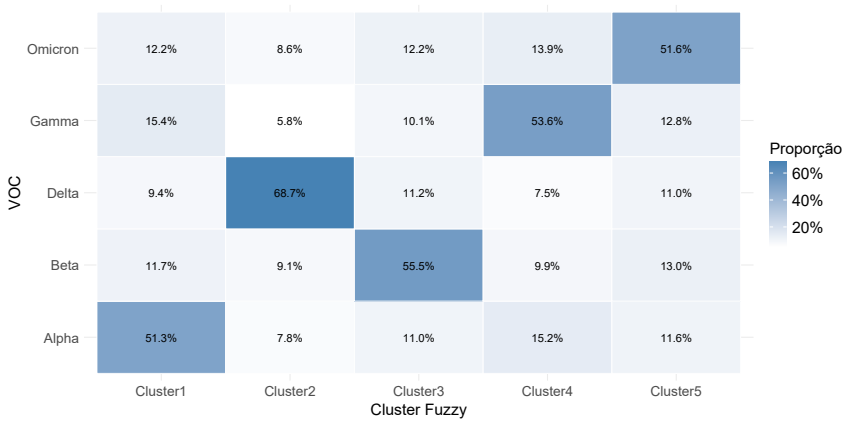


Figura 4.4: Heatmap da média de pertinência *fuzzy* de cada variante de preocupação (VOC) em relação aos clusters identificados, indicando o grau de associação entre variantes e agrupamentos.

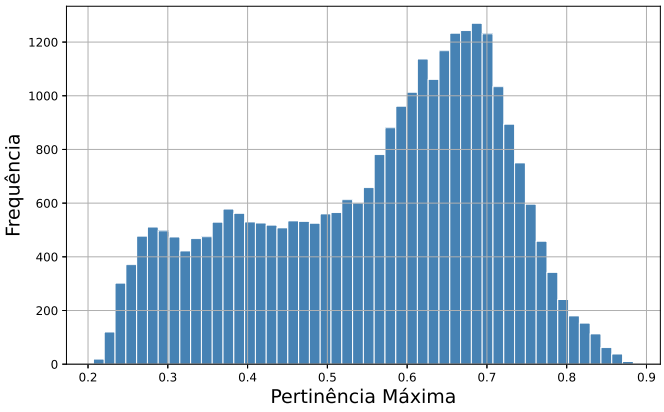
a todos os *clusters*. Valores de `PertinenciaMax` próximos a 1 indicam uma associação inequívoca com um único perfil mutacional, ao passo que valores inferiores sinalizam ambiguidade.

A distribuição dos valores de `PertinenciaMax` em todo o conjunto de dados é apresentada na Figura 4.5a. Observa-se que, embora uma parcela significativa das amostras exiba alta pertinência, uma fração considerável apresenta valores intermediários. Este fato sugere a existência de um subconjunto de amostras com características mistas, que não se encaixam perfeitamente em um único *cluster*.

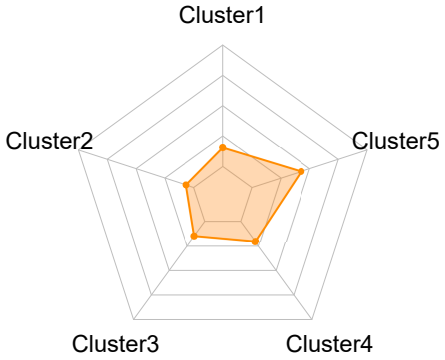
Com base nessa distribuição, definimos um limiar para identificar amostras de interesse, classificando aquelas com `PertinenciaMax` inferior a 0,6 como "altamente ambíguas". As Figuras 4.5b e 4.5c apresentam exemplos representativos dessas amostras por meio de gráficos de radar, que ilustram visualmente seu grau de associação com cada um dos cinco *clusters*. Essas representações gráficas evidenciam a existência de perfis transicionais, que compartilham características de múltiplos agrupamentos sem pertencer de forma clara a nenhum deles. Essa análise qualitativa reforça o potencial do modelo *fuzzy* para capturar a complexidade e a continuidade do espectro mutacional viral, revelando nuances que seriam perdidas em abordagens de *hard clustering*.

4.2.3 Determinação de Assinaturas Mutacionais por Exclusividade Fuzzy

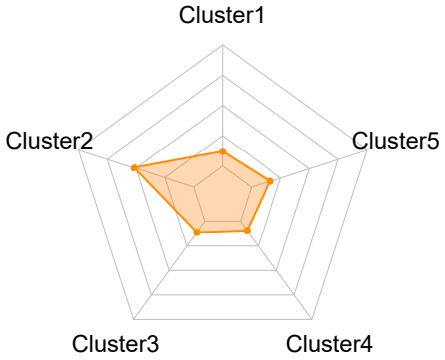
Após a identificação dos agrupamentos *fuzzy*, foi realizada uma investigação sobre quais mutações contribuem de forma mais específica para a definição de cada grupo. Para isso, foi conduzida uma análise de exclusividade mutacional, com o objetivo de identificar substituições de aminoácidos preferencialmente associadas a determinados *clusters*. Para cada mutação, calculou-se a soma ponderada de suas ocorrências pelo grau de pertinência das amostras em cada *cluster*, sendo esse valor normalizado de modo a refletir a proporção da chamada expressão *fuzzy* da mutação em relação aos diferentes grupos.



(a) Distribuição dos valores de `PertinenciaMax`.



(b) Amostra ambígua, exemplo 1.



(c) Amostra ambígua, exemplo 2.

Figura 4.5: Análise da ambiguidade *fuzzy* nas amostras, evidenciando perfis mutacionais com maior sobreposição entre clusters e regiões de transição no espaço vetorial.

Os resultados dessa análise são apresentados na Figura 4.6, que exibe, para cada *cluster*, as vinte mutações com maior grau de exclusividade, ordenadas pela proporção máxima de associação *fuzzy*. A inspeção visual desses gráficos evidencia dois padrões principais: em alguns agrupamentos, observam-se mutações altamente concentradas, com proporções superiores a 0.8, o que indica forte especificidade; em outros, as mutações são distribuídas de forma mais difusa entre os *clusters*, sugerindo compartilhamento de perfis mutacionais e possíveis zonas de convergência evolutiva. Esses padrões reforçam tanto a existência de assinaturas mutacionais distintas quanto a presença de regiões latentes no espaço genômico onde ocorrem interseções funcionais entre linhagens.

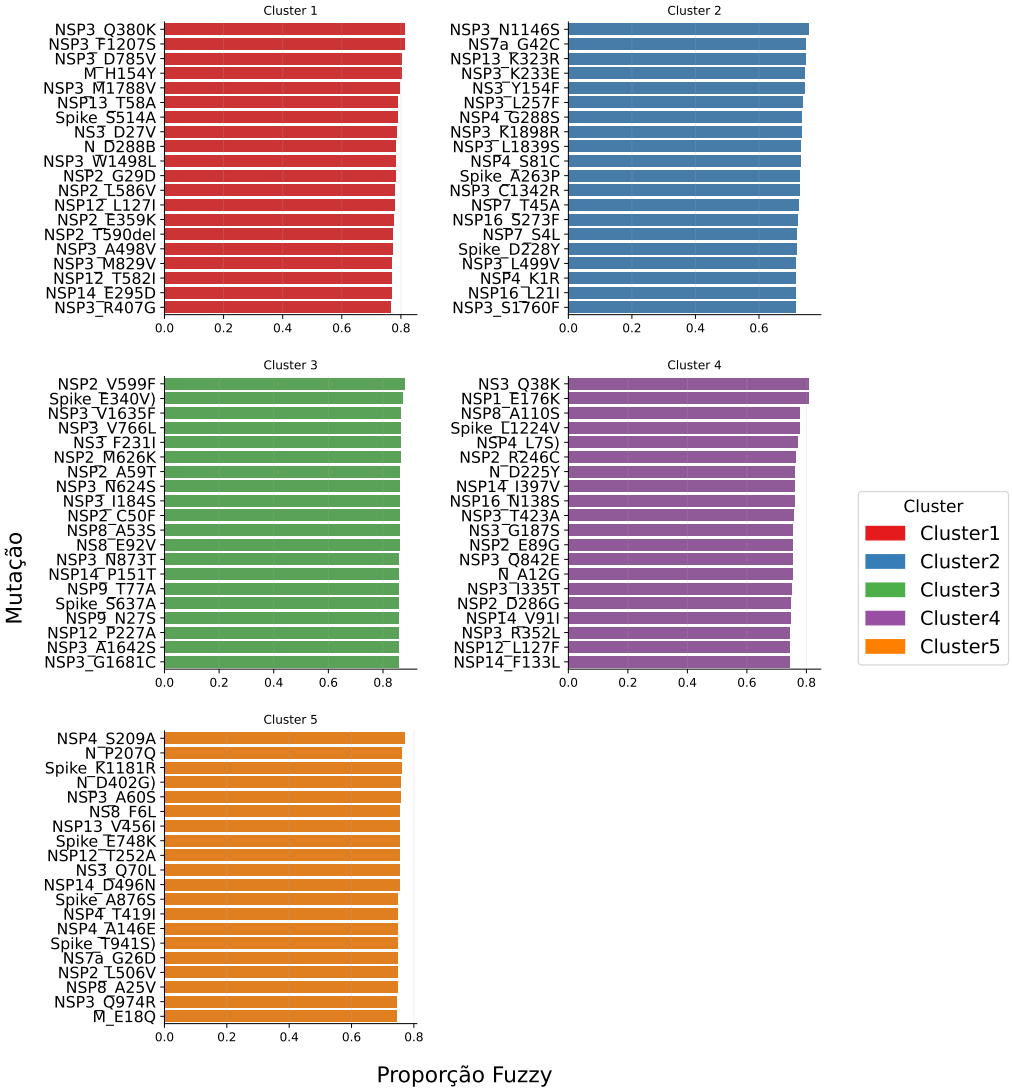


Figura 4.6: Proporção *fuzzy* das vinte mutações mais exclusivas para cada *cluster*. Cada barra representa a agregação ponderada da pertinência *fuzzy* das amostras em que a mutação ocorre, indicando seu grau de especificidade em relação aos demais agrupamentos.

A análise detalhada das mutações mais exclusivas permite a identificação de marcadores potencialmente relevantes do ponto de vista funcional e epidemiológico. No *Cluster*

1, destaca-se a substituição `NSP3_Q380K`, `NSP3_F1207S` com elevada concentração relativa. O *Cluster* 3 apresenta um perfil com exclusividade significativa para as mutações `NSP2_V599F`, `Spike_E340V`. Já o *Cluster* 2 é marcado pela ocorrência concentrada de `NSP3_N1146S` e `NS7a_G42C`, enquanto o *Cluster* 5 revela um conjunto de substituições exclusivas incluindo `NSP4_S209A`, `N_P207Q` e `Spike_K1181R`, e por fim o *Cluster* 4 que apresenta maior ocorrência nas mutações `NS3_Q38K` e `NSP1_E176K`.

A concentração dessas substituições em genes com funções críticas, como as proteínas não estruturais `NSP2`, `NSP3`, `NSP13` e `NSP14`, envolvidas na replicação e correção do RNA viral, bem como na proteína `Spike`, associada à entrada viral na célula hospedeira, conforme descrito na Seção 1.2, sugere que os agrupamentos *fuzzy* capturam não apenas correlações estatísticas, mas também estruturas latentes com significado biológico. Tais achados indicam que a clusterização *fuzzy*, ao integrar informação probabilística com perfis de mutações contextualmente representadas, é capaz de revelar padrões funcionais implícitos, reforçando seu potencial como ferramenta interpretável para vigilância genômica e análise evolutiva de variantes virais.

4.3 Resultados com Base no Desempenho de Modelos Supervisionados para Classificação de Variantes

Neste seção, discute-se o desempenho dos modelos de aprendizado supervisionado aplicados à tarefa de classificação das VOCs do SARS-CoV-2. A análise comparativa entre os classificadores `SVM`, `Random Forest`, `XGBoost` e `k-NN` permite não apenas identificar a abordagem mais eficaz, mas também fornece subsídios para compreender aspectos estruturais dos dados e a complexidade da tarefa de classificação a partir de *embeddings* genômicos.

Todos os classificadores foram treinados com os *embeddings* gerados a partir das substituições de aminoácidos, totalizando 382 variáveis após a remoção de duas dimensões com variância nula, algo já discutido na seção 3.3.4. A avaliação foi conduzida por meio de validação cruzada estratificada com 5 folds, permitindo uma estimativa confiável de capacidade preditiva dos modelos de classificação em questão.

4.3.1 As Limitações do SVM na Classificação de Variantes

O classificador `SVM`, com *kernel RBF*, foi testado com diferentes valores do parâmetro de penalização C (0,25, 0,50 e 1,00), mantendo-se fixo o valor de $\sigma = 0{,}00213$, que controla a largura do *kernel RBF*. Os resultados obtidos estão sumarizados na Tabela 4.2.

A melhor configuração foi obtida com $C = 0{,}25$, segundo o critério de menor *logLoss*, indicando maior confiabilidade probabilística nas predições. Nessa configuração, o modelo alcançou uma acurácia média de 93,31%, $F1$-score macro de 92,96%, *AUC* de 99,82%, *prAUC* de 99,33% e índice de *Kappa* de 91,63%. As métricas médias de sensibilidade (93,31%) e precisão (94,23%) também foram elevadas, enquanto a acurácia balanceada atingiu 95,82%, evidenciando um desempenho geral equilibrado na classificação das cinco variantes de preocupação, Alpha, Beta, Gamma, Delta e Omicron. A

Tabela 4.2: Resultados da validação cruzada para o classificador SVM com kernel RBF.

C	logLoss	AUC	prAUC	Acurácia	Kappa	F1 macro	Sens. média
0,25	0,4324	0,9982	0,9933	0,9331	0,9163	0,9296	0,9331
0,50	0,4384	0,9983	0,9936	0,9369	0,9211	0,9339	0,9369
1,00	0,4664	0,9980	0,9926	0,9380	0,9225	0,9351	0,9380

C	Espec. média	Prec. média	NPV média	Acc. balanceada
0,25	0,9833	0,9423	0,9844	0,9582
0,50	0,9842	0,9452	0,9853	0,9606
1,00	0,9845	0,9461	0,9855	0,9612

elevada *AUC* e *prAUC* refletem a capacidade discriminativa do modelo, mesmo diante de classes com padrões genômicos similares.

Embora os valores de $C = 0{,}5$ e $C = 1{,}0$ tenham proporcionado ligeiros aumentos em métricas como acurácia e $F1$-score, a escolha de $C = 0{,}25$ foi priorizada para maximizar a calibração das probabilidades preditas. Essa escolha é relevante em contextos biomédicos, onde a confiabilidade estatística das classificações pode ser tão importante quanto a taxa de acerto. A combinação do pré-processamento, com balanceamento estratificado das classes e a utilização de *embeddings* contextuais, permitiu que o `SVM` modelasse relações não lineares entre os perfis mutacionais das variantes.

A Figura 4.7 apresenta a matriz de confusão percentual obtida na validação cruzada. Observa-se que a maioria das variantes apresenta elevadas taxas de acerto, especialmente Beta e Alpha, que atingiram 99,9%. No entanto, a variante Gamma apresentou o maior índice de confusão relativa, com 22,7% de suas amostras sendo incorretamente classificadas como Omicron. Essa sobreposição pode ser explicada pela similaridade entre os perfis genômicos dessas duas variantes no espaço de *embeddings*. Esse padrão de erro persistiu mesmo com ajustes de hiperparâmetros, sugerindo uma limitação estrutural do modelo para separar classes com grande proximidade vetorial.

De forma geral, os resultados indicam que o `SVM` possui boa capacidade discriminativa, sendo capaz de separar corretamente a maioria das classes. Entretanto, seu desempenho foi inferior ao dos modelos baseados em árvores e ao `k-NN`, especialmente no que diz respeito à sensibilidade da variante Gamma. Tal comportamento pode ser atribuído à natureza contínua e suave das fronteiras de decisão do *kernel RBF*, que, embora eficientes para separar padrões simples, podem não capturar adequadamente estruturas latentes complexas e sobrepostas, como as observadas em contextos de alta similaridade genômica. Apesar dessas limitações, o modelo `SVM` demonstrou potencial como alternativa viável em cenários de classificação supervisionada em bioinformática, sobretudo quando interpretabilidade e estabilidade probabilística são desejáveis.

4.3.2 Modelos Baseados em Árvores de Decisão: XGBoost

Uma conclusão importante extraída da análise comparativa é a clara superioridade dos modelos baseados em *ensembles* de árvores de decisão, com destaque para o `XGBoost` e o `Random Forest`. Ambos os classificadores apresentaram desempenho excepcional, com métricas de acurácia, $F1$-score e *AUC* consistentemente superiores a $99{,}7\%$ durante a

validação cruzada. Esse resultado sugere que a estrutura hierárquica e não linear das árvores é muito eficaz na modelagem das complexas interações entre substituições de aminoácidos que caracterizam cada variante.

O classificador `XGBoost` destacou-se como o modelo de melhor desempenho geral, não apenas por apresentar métricas ligeiramente superiores em relação aos demais algoritmos avaliados, mas, sobretudo, por alcançar o menor valor de *logLoss*, métrica que avalia a calibração probabilística das predições. Em cenários biomédicos, esse indicador é relevante, pois expressa a confiança estatística nas classificações realizadas, um atributo essencial em sistemas de vigilância genômica. Além de ser preciso, o `XGBoost` demonstrou elevada confiabilidade ao atribuir probabilidades às suas predições.

Foram avaliadas diferentes combinações de hiperparâmetros, variando-se o número total de iterações (`nrounds`), a profundidade máxima das árvores (`max depth`) e a fração de amostras usadas por árvore (`subsample`). Os demais parâmetros, como taxa de aprendizado (`eta`) e proporção de colunas por árvore (`colsample_bytree`), foram mantidos fixos nos valores padrão de implementação, 0,3 e 0,6, respectivamente. A melhor configuração encontrada foi composta por `nrounds` = 150, `max depth` = 2, `subsample` = 0,5, cujos resultados detalhados estão apresentados na Tabela 4.3.

Tabela 4.3: Melhores configurações do XGBoost e suas métricas de desempenho.

nrounds	**max_depth**	**subsample**	**logLoss**	**Acurácia**	**F1 macro**
150	2	0,5	0,0068	0,9983	0,9983
100	2	0,5	0,0069	0,9982	0,9982
150	1	0,5	0,0069	0,9982	0,9982

nrounds	**max_depth**	**subsample**	**AUC**	**prAUC**
150	2	0,5	0,99998	0,9883
100	2	0,5	0,99998	0,9963
150	1	0,5	0,99998	0,9971

Com essa parametrização, o modelo obteve um *logLoss* de apenas 0,0068, indicando altíssima confiança nas probabilidades preditas. A acurácia e o $F1$-score macro atingiram 99,83%, refletindo equilíbrio entre sensibilidade e precisão. A métrica *AUC* de 0,99998 confirmou a excelente capacidade discriminativa do modelo, mesmo diante de classes com padrões mutacionais semelhantes, enquanto a *prAUC* de 0,9883 reforçou sua robustez sob diferentes limiares de decisão. O índice de *Kappa* com 0,9977 e a acurácia balanceada com 99,88% corroboram o desempenho consistente do `XGBoost` em todo o espectro de classes avaliadas.

Comparado aos demais classificadores, o `XGBoost` apresentou desempenho superior tanto em termos de calibração probabilística quanto em métricas tradicionais de classificação. Sua estabilidade, mesmo com profundidade de árvore reduzida e taxas moderadas de subamostragem, evidencia seu potencial para aplicações de classificação genômica em larga escala. A Figura 4.8 apresenta a matriz de confusão percentual do modelo, na qual se observa desempenho excepcional em todas as classes: 99,9% de acerto para Alpha, Beta e Gamma, 99,8% em Delta e 99,6% para Omicron. As taxas de confusão foram mínimas, por exemplo, Omicron teve 0,3% de amostras incorretamente classificadas como

Delta, indicando alta separabilidade mesmo entre variantes com perfis genômicos similares. Essa distribuição concentrada de erros fora da diagonal confirma visualmente os resultados quantitativos e reforça a confiabilidade do modelo.

4.3.3 Desempenho do Modelo Random Forest

O classificador `Random Forest` também apresentou desempenho quase impecável na tarefa de classificação das variantes do SARS-CoV-2. No processo de otimização, foi ajustado o hiperparâmetro `mtry`, que define o número de variáveis, ou atributos, candidatas a serem avaliadas em cada divisão das arvores de decisão, influenciando diretamente a diversidade e a robustez do conjunto de árvores. Foram testadas três valores, 2, 192 e 382, e a melhor configuração foi obtida com `mtry` = 192, conforme o critério indicado pelo menor valoe de *logLoss*.

Nessa configuração, o modelo alcançou acurácia de 99,71%, $F1$-score macro de 99,71%, *AUC* de 0,9999 e *logLoss* de 0,0290, conforme detalhado na Tabela 4.4. As métricas complementares também evidenciaram desempenho notável, a *prAUC* foi de 0,2305, refletindo a influência do balanceamento entre classes na distribuição de precisão e sensibilidade, enquanto o índice de *Kappa* atingiu 0,9964, indicando forte concordância entre predições e rótulos reais. O equilíbrio entre precisão (99,71%), sensibilidade (99,71%) e acurácia balanceada (99,82%) confirma a robustez do modelo em múltiplas dimensões avaliadas.

Tabela 4.4: Resultados da validação cruzada para o classificador Random Forest.

mtry	logLoss	AUC	prAUC	Acurácia	Kappa	F1 macro	Sens. média
2	0,0673	0,9999	0,6560	0,9964	0,9955	0,9964	0,9964
192	0,0290	0,9999	0,2305	0,9971	0,9964	0,9971	0,9971
382	0,0319	0,9999	0,1668	0,9946	0,9932	0,9946	0,9946

mtry	Espec. média	Prec. média	NPV média	Acc. balanceada
2	0,9991	0,9964	0,9991	0,9977
192	0,9993	0,9971	0,9993	0,9982
382	0,9986	0,9946	0,9986	0,9966

A eficácia do `Random Forest` pode ser atribuída à sua capacidade intrínseca de mitigar o sobreajuste por meio da agregação de múltiplas árvores de decisão construídas a partir de subconjuntos aleatórios de amostras e variáveis. Essa abordagem gera uma diversidade estrutural entre as árvores, favorecendo a generalização sem perda de capacidade discriminativa. A Figura 4.9 apresenta a matriz de confusão percentual obtida durante a validação cruzada. Observa-se que as classes Alpha, Beta, Delta e Gamma foram corretamente classificadas em 99,8% dos casos. No caso da variante Delta, por exemplo, houve apenas 0,2% de confusão com a variante Omicron. A variante Omicron, embora tenha apresentado uma distribuição de erros um pouco mais dispersa, ainda alcançou apresentou 99,4% de acertos, com desvios residuais mínimos entre as demais classes.

Em conjunto, esses resultados demonstram que o `Random Forest` não apenas é altamente eficaz na separação entre perfis mutacionais distintos, como também apresenta



Figura 4.7: Matriz de confusão percentual para o classificador SVM com kernel radial, obtida durante a validação cruzada estratificada (5-fold), ilustrando o desempenho do modelo na distinção entre classes.

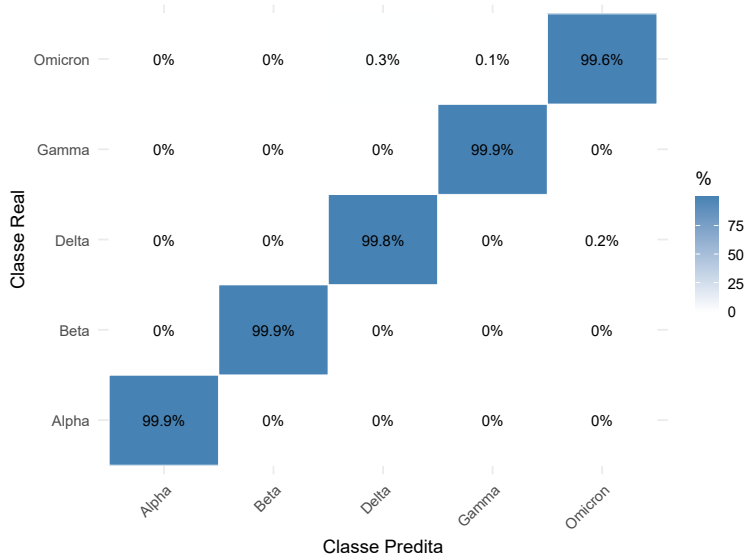


Figura 4.8: Matriz de confusão percentual para o classificador XGBoost, obtida durante a validação cruzada estratificada (5-fold), evidenciando a elevada taxa de acerto e a baixa confusão entre classes.

confiabilidade probabilística adequada para aplicações em vigilância genômica. As matrizes de confusão tanto do `Random Forest` quanto do `XGBoost` reforçam visualmente essa conclusão, exibindo margens de erro quase nulas mesmo diante da complexidade genômica subjacente.

4.3.4 O Desempenho do k-NN e suas Implicações

Outro achado relevante foi o desempenho altamente competitivo do classificador `k-NN`. Apesar de sua simplicidade conceitual, o modelo obteve resultados comparáveis aos dos algoritmos baseados em árvores de decisão como mostrada na Tabela 4.5, superando significativamente o desempenho observado para o `SVM`.

Tabela 4.5: Resultados da validação cruzada para o classificador k-NN.

k	logLoss	AUC	prAUC	Acurácia	Kappa	F1 macro	Sens. média
5	0,0372	0,9994	0,0074	0,9971	0,9964	0,9971	0,9971
7	0,0322	0,9995	0,0098	0,9969	0,9961	0,9969	0,9969
9	0,0316	0,9995	0,0118	0,9968	0,9960	0,9968	0,9968
11	0,0310	0,9996	0,0134	0,9968	0,9960	0,9968	0,9968
13	0,0303	0,9996	0,0149	0,9966	0,9957	0,9966	0,9966

k	Espec. média	Prec. média	NPV média	Acc. balanceada
5	0,9993	0,9971	0,9993	0,9982
7	0,9992	0,9969	0,9992	0,9980
9	0,9992	0,9968	0,9992	0,9980
11	0,9992	0,9968	0,9992	0,9980
13	0,9991	0,9966	0,9991	0,9978

Foram testados cinco valores do hiperparâmetro k (5, 7, 9, 11 e 13), sendo a configuração com $k = 13$ a que obteve o menor valor de *logLoss* (0,0303), indicando sua superioridade segundo o critério probabilístico. Nessa configuração, o modelo alcançou acurácia e $F1$-score macro de 99,66%, *AUC* de 0,9996, *prAUC* de 0,0149 e índice de *Kappa* de 0,9957. As métricas médias de precisão, sensibilidade, especificidade e acurácia balanceada também foram elevadas, todas acima de 99,6%, com destaque para a acurácia balanceada de 99,78%.

Embora o valor de *logLoss* tenha sido inferior ao observado no `XGBoost`, o desempenho global do `k-NN` demonstrou-se notavelmente estável e competitivo, mesmo operando em um espaço vetorial de alta dimensionalidade. Isso sugere que os *embeddings* gerados capturaram de maneira eficaz as similaridades genômicas entre as amostras. Para que um classificador baseado exclusivamente em distância funcione bem em um espaço de 382 dimensões, é necessário que os dados de cada classe formem regiões compactas e bem separadas. Assim, a alta performance do `k-NN` atua como uma validação indireta da qualidade das representações vetoriais, os *embeddings* foram capazes de mapear padrões mutacionais em uma geometria latente coerente.

A Figura 4.10 apresenta a matriz de confusão percentual obtida durante a validação cruzada. Observa-se que as classes Alpha, Beta, Delta e Gamma atingiram mais de 99,8% de acurácia, com taxas de erro praticamente nulas. A variante Alpha, por exemplo, foi

corretamente classificada em 99,9% dos casos, com apenas 0,1% de confusão para Omicron. Já a classe Omicron apresentou 98,9% de acerto, com 0,8% de suas amostras sendo equivocadamente classificadas como Delta. Apesar dessas pequenas taxas de confusão, o desempenho geral do `k-NN` manteve-se elevado, confirmando sua eficácia mesmo em cenários de alta similaridade genômica entre variantes.

Esses resultados demonstram que, mesmo sendo um algoritmo simples, o `k-NN` é capaz de produzir classificações precisas e confiáveis quando apoiado por representações vetoriais bem estruturadas. Sua utilização prática é facilitada por sua natureza não-paramétrica e interpretável, o que o torna uma alternativa viável e eficiente para tarefas de classificação em contextos biomoleculares complexos.

4.4 Desempenho Comparativo dos classificadores e Validação em Dados Externos

Uma vez visto os resultados dos quatro modelos aplicados na pesquisa, a Tabela 4.6 apresenta um resumo comparativo do desempenho dos classificadores. Dentre os modelos analisados, o `XGBoost` destacou-se em todas as principais métricas, pois alcançou a maior acurácia (99,83%), $F1$-score macro (99,83%), *AUC* (0,99998), índice de $Kappa$ (0,9977) e acurácia balanceada (99,88%). Além disso, obteve o menor valor de *logLoss* (0,0068), indicando excelente calibragem probabilística e elevada confiança nas predições.

Tabela 4.6: Comparativo entre os classificadores com base na validação cruzada (5-fold).

Modelo	Acurácia	F1 macro	logLoss	AUC	Kappa	Acc. balanceada
SVM	0,9331	0,9296	0,4324	0,9982	0,9163	0,9582
Random Forest	0,9971	0,9971	0,0290	0,9999	0,9964	0,9982
XGBoost	**0,9983**	**0,9983**	**0,0068**	**0,99998**	**0,9977**	**0,9988**
k-NN	0,9966	0,9966	0,0303	0,9996	0,9957	0,9978

Os modelos `Random Forest` e `k-NN` também apresentaram desempenhos altamente competitivos, com acurácia acima de 99,6%, valores de $Kappa$ superiores a 0,995 e *AUC* próximas de 1. O `Random Forest` apresentou resultados ligeiramente superiores ao `k-NN`, especialmente nas métricas *logLoss* e *prAUC*. Já o classificador `SVM`, embora tenha atingido desempenho satisfatório, obteve métricas inferiores às dos demais modelos, com acurácia de 93,31% e *logLoss* de 0,4324, refletindo menor confiança probabilística.

Esses resultados indicam que os modelos baseados em árvores de decisão, `XGBoost` e `Random Forest`, são mais eficazes para a tarefa de classificação de variantes do SARS-CoV-2 a partir de *embeddings* de substituições de aminoácidos. Tais modelos se mostraram não apenas mais precisos, mas também mais robustos e generalizáveis, mesmo em um espaço vetorial de alta dimensionalidade.

A Tabela 4.7 apresenta os resultados obtidos na avaliação dos modelos sobre o conjunto de teste externo, composto por amostras que não participaram do treinamento nem da validação cruzada. De modo geral, todos os modelos mantiveram altos níveis de acurácia, demonstrando boa capacidade de generalização. O `XGBoost` novamente se destacou,

obtendo as maiores acurácias em todas as variantes, com 99,91% para Alpha, 99,83% para Delta, 99,75% para Gamma e 99,61% para Omicron.

Tabela 4.7: Acurácia por modelo e por VOC no conjunto de teste externo.

Modelo	VOC	Acurácia	Total de Amostras
SVM	Alpha	0,9988	45.099
RF		0,9978	
XGB		**0,9991**	
KNN		0,9988	
SVM	Delta	0,9962	223.466
RF		0,9971	
XGB		**0,9983**	
KNN		0,9976	
SVM	Gamma	0,6677	6.937
RF		0,9967	
XGB		0,9975	
KNN		**0,9977**	
SVM	Omicron	0,9892	12.492
RF		0,9922	
XGB		**0,9961**	
KNN		0,9894	

Os modelos `Random Forest` e `k-NN` apresentaram desempenhos muito próximos ao do `XGBoost`, com acurácia superior a 99% em todas as classes. O `SVM`, por sua vez, mostrou desempenho comparável nas variantes Alpha, Delta e Omicron, mas apresentou acurácia significativamente inferior na variante Gamma (66,77%), o que confirma a tendência observada na validação cruzada quanto à sua dificuldade em separar essa classe.

A análise final, baseada na avaliação com dados não vistos durante o treinamento, reforça a eficiência e confiabilidade dos modelos de melhor desempenho. O fato de `XGBoost`, `Random Forest` e `k-NN` manterem sua alta performance em dados totalmente novos demonstra que os padrões aprendidos não são artefatos do conjunto de treinamento, mas sim representações genuínas das assinaturas mutacionais das variantes.

Em síntese, os resultados apresentados demonstram que a combinação entre *embeddings* de substituições de aminoácidos e modelos supervisionados, principalmente aqueles baseados em árvores de decisão, constitui uma abordagem poderosa para a classificação automática de variantes do SARS-CoV-2. Essa estratégia não apenas oferece elevada precisão, como também proporciona predições confiáveis e interpretáveis, características essenciais para sua adoção em sistemas reais de vigilância genômica. A implementação dessa metodologia pode acelerar a detecção de linhagens emergentes, subsidiando ações de saúde pública mais ágeis e baseadas em evidências genômicas.



Figura 4.9: Matriz de confusão percentual para o classificador Random Forest,obtida durante a validação cruzada estratficada (5-fold), destacando o desempenho do modelo na classificação das variantes.

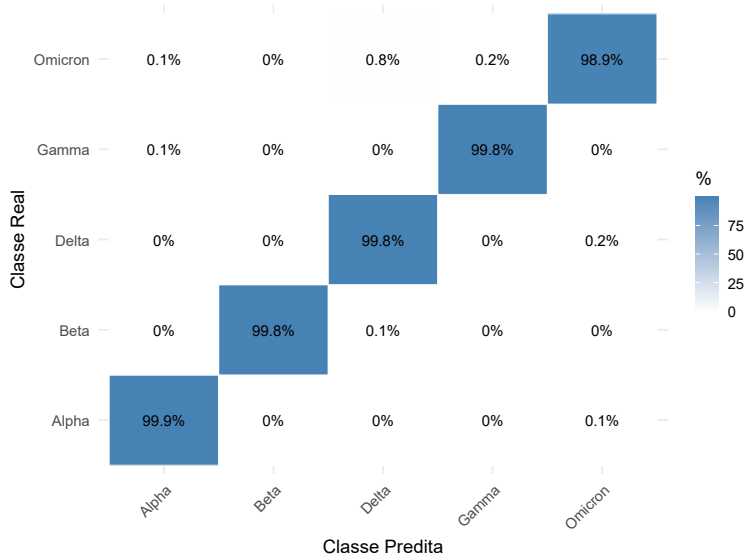


Figura 4.10: Matriz de confusão percentual para o classificador k-NN, obtida durante a validação cruzada estratificada (5-fold), apresentando a distribuição dos acertos e erros entre as classes avaliadas.

Capítulo 5

Conclusão

Este trabalho apresentou um conjunto de abordagens computacionais para a análise e classificação de variantes do SARS-CoV-2, baseadas em representações vetoriais derivadas de *AASubstitutions* e técnicas de aprendizado de máquina. Os experimentos realizados ao longo da pesquisa exploraram tanto métodos não supervisionados, como `Fuzzy C-Means` e `K-means`, quanto algoritmos supervisionados clássicos, como `XGBoost`, `Random Forest`, `SVM` e `k-NN`. Em todas as frentes, a utilização de *embeddings* contextuais, gerados por modelos baseados na arquitetura *Transformer*, demonstrou ser uma estratégia eficiente, escalável e altamente interpretável para a vigilância genômica viral.

As abordagens não supervisionadas permitiram identificar estruturas latentes nos dados genômicos, revelando agrupamentos coerentes com as VOCs, mesmo sem a utilização de rótulos durante o treinamento. A aplicação do modelo `Fuzzy C-Means`, destacou-se por capturar padrões de transição e ambiguidade entre variantes, fornecendo *insights* sobre regiões de sobreposição e exclusividade mutacional. A análise *fuzzy* evidenciou a presença de assinaturas específicas em genes funcionalmente relevantes, como `NSP12`, `NSP13` e `Spike`, e revelou amostras com perfis híbridos ou transicionais, importantes para a compreensão da evolução viral.

Já as abordagens supervisionadas, com destaque para os modelos baseados em árvores de decisão, apresentaram desempenho excepcional na tarefa de classificação de variantes. O `XGBoost` obteve os melhores resultados tanto na validação cruzada quanto em um conjunto externo de teste composto por quase 288 mil amostras, alcançando acurácia de $99,83\%$ e *logLoss* de apenas $0,0068$. Esses resultados indicam forte capacidade de generalização e reforçam o potencial do método para aplicações reais em cenários de vigilância automatizada. Importante ressaltar que o uso exclusivo das substituições de aminoácidos como entrada, sem necessidade de alinhamento genômico ou metadados clínicos, simplifica o pipeline computacional e amplia sua aplicabilidade.

De forma geral, a combinação entre *embeddings* de *AASubstitutions* e algoritmos de aprendizado de máquina revelou-se eficaz para representar e discriminar a diversidade mutacional das variantes virais. O trabalho contribui metodologicamente ao propor uma abordagem alternativa às representações tradicionais baseadas em nucleotídeos ou sequências proteicas completas, focando diretamente no impacto funcional das mutações. Além disso, a estratégia adotada abre caminho para a incorporação de representações vetoriais em sistemas inteligentes, como arquiteturas do tipo *Retrieval-Augmented Generation* (RAG), otimizando a integração entre vigilância genômica e inteligência artificial.

5.1 Trabalhos Futuros

Dentre as perspectivas futuras, propõe-se expandir a metodologia desenvolvida, explorando novas possibilidades de análise e avaliação dos embeddings gerados a partir de substituição de aminoácidos. Uma primeira linha de investigação consiste na investigação de novos algoritmos de clusterização, especialmente aqueles capazes de identificar padrões em distribuições complexas e com diferentes densidades. Métodos baseados em densidade, como o *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN), e sua extensão hierárquica *Hierarchical DBSCAN* (HDBSCAN), permitem detectar agrupamentos com formatos arbitrários, além de manejar ruídos e regiões de transição de forma mais natural do que abordagens particionais. Tais métodos podem oferecer vantagens importantes na caracterização de linhagens emergentes, raras ou estruturalmente ambíguas.

Outra possibilidade de expansão refere-se à adoção de técnicas alternativas de redução de dimensionalidade mais adequadas à estrutura não linear e à alta dimensionalidade dos embeddings. O *Uniform Manifold Approximation and Projection* (UMAP) já é reconhecido por preservar melhor as relações topológicas globais e locais dos dados do que o t-SNE, sendo uma opção promissora para visualizações mais estáveis e interpretáveis. Além disso, métodos baseados em aprendizado profundo, como *autoencoders*, oferecem uma abordagem treinável e adaptativa para compressão não linear dos *embeddings*, podendo ser utilizados tanto para visualização quanto para pré-processamento antes da aplicação de algoritmos supervisionados ou não supervisionados.

No contexto da classificação supervisionada, futuras investigações podem concentrar-se na avaliação da generalização da metodologia por meio da realização de testes com conjuntos de dados provenientes de outras bases genômicas, além do GISAID. Essa ampliação permitiria analisar a robustez, a transferibilidade e o comportamento dos classificadores em cenários distintos de coleta, curadoria e anotação dos dados, contribuindo para uma compreensão mais ampla das limitações e potencialidades da abordagem proposta.

Por fim, destaca-se o potencial dos *embeddings* derivados das *AASubstitutions* como base para sistemas de recuperação de informação RAG. Nessa perspectiva, cada vetor representa uma amostra viral codificada semanticamente com base em seu perfil mutacional, e pode ser indexado em bancos vetoriais como `ChromaDB`. A recuperação por similaridade vetorial permite a construção de consultas semânticas eficientes, em que modelos gerativos podem ser conectados a esse banco para responder perguntas sobre linhagens, mutações específicas, regiões geográficas ou propriedades funcionais. Essa integração entre representação distribuída e geração neural oferece uma rota promissora para o desenvolvimento de sistemas interativos e explicáveis de vigilância genômica assistida por inteligência artificial.

Referências Bibliográficas

Aggarwal, Charu C., Alexander Hinneburg e Daniel A. Keim ,(2001), On the surprising behavior of distance metrics in high dimensional space, *em* 'International Conference on Database Theory', Springer, pp. 420–434.
URL: *https://link.springer.com/chapter/10.1007/3-540-44503-X_27*

Ali, Rechfy Kasem Abou ,(2024), Estudo dos polimorfismos nos genes das citocinas IL1, IL6, IL18 e do inflamassoma NLRP1 e NLRP3 em pacientes convalescentes da COVID-19 e pacientes com diagnóstico de COVID-19 no estado do Amazonas, Brasil, Tese de doutorado, Universidade Federal do Amazonas, Instituto de Ciências Biológicas, Programa de Pós-Graduação em Imunologia Básica e Aplicada, Manaus. Orientadora: Dra. Adriana Malheiro Alle Marie; Coorientadora: Dra. Andréa Monteiro Tarragô.
URL: *https://tede.ufam.edu.br/bitstream/tede/10139/2/Tese_RechfyAbouAli_PPGIBA.pdf*

Ali, S., B. Sahoo, N. Ullah, A. Zelikovskiy, M. Patterson e I. Khan ,(2021), A k-mer based approach for sars-cov-2 variant identification, *em* 'Bioinformatics Research and Applications: 17th International Symposium, ISBRA 2021, Shenzhen, China, November 26–28, 2021, Proceedings', Vol. 13084 de *Lecture Notes in Computer Science*, Springer, pp. 153–164.
URL: *https://arxiv.org/abs/2108.03465*

Andersen, Kristian G., Andrew Rambaut, W. Ian Lipkin, Edward C. Holmes e Robert F. Garry ,(2020), 'The proximal origin of sars-cov-2', *Nature Medicine* **26**(4), 450–452.
URL: *https://doi.org/10.1038/s41591-020-0820-9*

Azevedo, K. S., L. C. de Souza, M. G. F. Coutinho, R. d. M. Barbosa e M. A. C. Fernandes ,(2024), 'Deepvirusclassifier: A deep learning tool for classifying sars-cov-2 based on viral subtypes within the coronaviridae family', *BMC Bioinformatics* **25**(231).
URL: *https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-024-05754-1*

Bai, Chongzhi, Qiming Zhong e George Fu Gao ,(2022), 'Overview of sars-cov-2 genome-encoded proteins', *Science China Life Sciences* **65**(2), 280–294.
URL: *https://doi.org/10.1007/s11427-021-1964-4*

Beduk, D., J. I. de Oliveira Filho, T. Beduk, D. Harmanci, F. Zihnioglu, C. Cicek, R. Sertoz, B. Arda, T. Goksel, K. Turhan et al. ,(2022), ''all in one' sars-cov-2 variant

recognition platform: Machine learning-enabled point of care diagnostics', *Biosensors and Bioelectronics: X* **10**, 100105.
URL: *https://doi.org/10.1016/j.biosx.2022.100105*

Beltagy, Iz, Arman Cohan e Kyle Lo ,(2019), 'Scibert: Pretrained contextualized embeddings for scientific text', *CoRR* **abs/1903.10676**.
URL: *http://arxiv.org/abs/1903.10676*

Bezdek, James C. ,(1981), *Pattern Recognition with Fuzzy Objective Function Algorithms*, Springer Science+Business Media.
URL: *https://doi.org/10.1007/978-1-4757-0450-1*

Bhattacharya, Manojit, Ashish Ranjan Sharma, Kuldeep Dhama, Govindasamy Agoramoorthy e Chiranjib Chakraborty ,(2022), 'Omicron variant (b.1.1.529) of sars-cov-2: understanding mutations in the genome, s-glycoprotein, and antibody-binding regions', *GeroScience* **44**(2), 619–637.
URL: *https://doi.org/10.1007/s11357-022-00532-4*

Breiman, Leo ,(2001), 'Random forests', *Machine Learning* **45**(1), 5–32.
URL: *https://doi.org/10.1023/A:1010933404324*

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever e Dario Amodei ,(2020), 'Language models are few-shot learners'.
URL: *https://arxiv.org/abs/2005.14165*

Calinski, T. e J. Harabasz ,(1974), 'A dendrite method for cluster analysis', *Communications in Statistics* **3**(1), 1–27.
URL: *https://www.tandfonline.com/doi/abs/10.1080/03610927408827101*

Camacho-Collados, Jose e Mohammad Taher Pilehvar ,(2020), Embeddings in natural language processing, *em* L.Specia e D.Beck, eds., 'Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts', International Committee for Computational Linguistics, Barcelona, Spain (Online), pp. 10–15.
URL: *https://aclanthology.org/2020.coling-tutorials.2/*

Câmara, G. B. M., M. G. F. Coutinho, L. M. D. d. Silva, W. V. d. N. Gadelha, M. F. Torquato, R. d. M. Barbosa e M. A. C. Fernandes ,(2022), 'Convolutional neural network applied to sars-cov-2 sequence classification', *Sensors* **22**(15), 5730.
URL: *https://www.mdpi.com/1424-8220/22/15/5730*

Chen, Tianqi e Carlos Guestrin ,(2016), Xgboost: A scalable tree boosting system, *em* 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge

Discovery and Data Mining', KDD '16, Association for Computing Machinery, New York, NY, USA, p. 785–794.
URL: *https://doi.org/10.1145/2939672.2939785*

Cheng, Peng, Jianyi Yang, Dapeng Xiong e Yang Zhang ,(2024), 'Zero-shot prediction of mutation effects with multimodal deep representation learning guides protein engineering', *Cell Research* **34**(5), 401–415.
URL: *https://www.nature.com/articles/s41422-024-00902-w*

Cherian, Sarah, Varsha Potdar, Santosh Jadhav, Pragya Yadav, Nivedita Gupta, Manohar Das, Pradeep Rakshit, Sreenivas Singh, Priya Abraham, Samiran Panda e National Covid Genomics Team ,(2021), 'Sars-cov-2 spike mutations, l452r, t478k, e484q and p681r, in the second wave of covid-19 in maharashtra, india', *Microorganisms* **9**(7), 1542.
URL: *https://www.mdpi.com/2076-2607/9/7/1542*

Chourasia, P., T. Murad, Z. Tayebi, S. Ali, I. U. Khan e M. Patterson ,(2023), Efficient classification of sars-cov-2 spike sequences using federated learning, *em* 'Annual International Conference on Information Management and Big Data (SimBig)', Springer, pp. 80–96.
URL: *https://arxiv.org/abs/2302.08688*

Cortes, Corinna e Vladimir Vapnik ,(1995), 'Support-vector networks', *Machine Learning* **20**(3), 273–297.
URL: *https://doi.org/10.1007/BF00994018*

Coutinho, M. G. F., G. B. M. Câmara, R. d. M. Barbosa e M. A. C. Fernandes ,(2023), 'Sars-cov-2 virus classification based on stacked sparse autoencoder', *Computational and Structural Biotechnology Journal* **21**, 284–298.
URL: *https://doi.org/10.1016/j.csbj.2022.12.007*

Cover, T. e P. Hart ,(1967), 'Nearest neighbor pattern classification', *IEEE Transactions on Information Theory* **13**(1), 21–27.
URL: *https://ieeexplore.ieee.org/document/1053964*

Câmara, Gabriel Bezerra Motta ,(2024), Advanced Convolutional Neural Network Techniques for Classification of SARS-CoV-2 Variants and Other Viruses: A Study Using k-mers and Chaos Game Representation, Tese (doutorado em bioinformática), Universidade Federal do Rio Grande do Norte, Instituto Metrópole Digital, Natal, RN, Brasil. Orientador: Prof. Dr. Marcelo Augusto Costa Fernandes.
URL: *https://repositorio.ufrn.br/handle/123456789/61974*

Davies, David L. e Donald W. Bouldin ,(1979), 'A cluster separation measure', *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2), 224–227.
URL: *https://doi.org/10.1109/TPAMI.1979.4766909*

de Souza, L. C., K. S. Azevedo, J. G. de Souza, R. d. M. Barbosa e M. A. C. Fernandes ,(2023), 'New proposal of viral genome representation applied in the classification of sars-cov-2 with deep learning', *BMC Bioinformatics* **24**(92).
URL: *https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-023-05188-1*

Deb, Pragyan, Davide Furceri, Jonathan D. Ostry e Nour Tawk ,(2022), 'The economic effects of covid-19 containment measures', *Open Economies Review* **33**(1), 1–32.
URL: *https://doi.org/10.1007/s11079-021-09638-2*

Devlin, Jacob, Ming-Wei Chang, Kenton Lee e Kristina Toutanova ,(2018), 'BERT: pre-training of deep bidirectional transformers for language understanding', *CoRR* **abs/1810.04805**.
URL: *http://arxiv.org/abs/1810.04805*

Dunn, J. C. ,(1973), 'A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters', *Journal of Cybernetics* **3**(3), 32–57.
URL: *https://doi.org/10.1080/01969727308546046*

Elnaggar, Ahmed, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik e Burkhard Rost ,(2022), 'Prottrans: Toward understanding the language of life through self-supervised learning', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(10), 7112–7127.
URL: *https://doi.org/10.1109/TPAMI.2021.3095381*

Fang, Letian, Jie Xu, Yue Zhao, Junyan Fan, Jiaying Shen, Wenbin Liu e Guangwen Cao ,(2023), 'The effects of amino acid substitution of spike protein and genomic recombination on the evolution of sars-cov-2', *Frontiers in Microbiology* **14**, 1228128. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
URL: *https://www.frontiersin.org/articles/10.3389/fmicb.2023.1228128/full*

Farhadpour, Sarah, Timothy A. Warner e Aaron E. Maxwell ,(2024), 'Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: Guidance and best practices', *Remote Sensing* **16**(3).
URL: *https://www.mdpi.com/2072-4292/16/3/533*

Faria, Nuno R, Thomas A Mellan, Charles Whittaker, Ingra M Claro, Darlan da S Candido, Swapnil Mishra, Myuki A E Crispim, Fernanda C Sales, Izabella Hawryluk, John T McCrone et al. ,(2021), 'Genomics and epidemiology of the p.1 sars-cov-2 lineage in manaus, brazil', *Science* **372**(6544), 815–821.
URL: *https://www.science.org/doi/10.1126/science.abh2644*

Fatima, N. e A. Ahmad ,(2024), 'Sars-cov-2 virus variant detection and mortality prediction through symptom analysis using machine learning', *Engineering Applications of Artificial Intelligence* **130**, 107743.
URL: *https://www.sciencedirect.com/science/article/pii/S0952197623019279*

Fawcett, Tom ,(2006), 'An introduction to roc analysis', *Pattern Recognition Letters* **27**(8), 861–874.
URL: *https://www.sciencedirect.com/science/article/pii/S016786550500303X*

Grandini, Margherita, Enrico Bagli e Giorgio Visani ,(2020), 'Metrics for multi-class classification: an overview'.
URL: *https://arxiv.org/abs/2008.05756*

Hanley, James A. e Barbara J. McNeil ,(1982), 'The meaning and use of the area under a receiver operating characteristic (roc) curve', *Radiology* **143**(1), 29–36.

Hartigan, J. A. e M. A. Wong ,(1979), 'Algorithm as 136: A k-means clustering algorithm', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28**(1), 100–108.
URL: *https://doi.org/10.2307/2346830*

Huang, C., Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Z. Cheng, T. Yu, J. Xia, Y. Wei, W. Wu, X. Xie, W. Yin, H. Li, M. Liu, Y. Xiao e B. Cao ,(2020), 'Clinical features of patients infected with 2019 novel coronavirus in wuhan, china', *The Lancet* **395**(10223), 497–506.
URL: *https://doi.org/10.1016/S0140-6736(20)30183-5*

Hugging Face ,(2024), 'all-minilm-l6-v2'. Accessed: 2024-10-30.
URL: *https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2*

Jolliffe, Ian T. e Jorge Cadima ,(2016), 'Principal component analysis: a review and recent developments', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374**(2065), 20150202.
URL: *http://doi.org/10.1098/rsta.2015.0202*

Jones, Bruce e Gregory Dixon ,(2024), 'Boosting textual understanding in llms with context-aware flexible length tokenization', *OSF Preprints* .
URL: *https://doi.org/10.31219/osf.io/9gnjt*

Khare, S., C. Gurry, L. Freitas, M. B. Schultz, G. Bach, A. Diallo, N. Akite, J. Ho, R. T. Lee, N. K. Yeo, Curation Team e S. Maurer-Stroh ,(2021), 'Gisaid's role in pandemic response', *China CDC Weekly* **3**(49), 1049–1051.
URL: *https://doi.org/10.46234/ccdcw2021.255*

Kourou, Konstantina, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis e Dimitrios I. Fotiadis ,(2015), 'Machine learning applications in cancer prognosis and prediction', *Computational and Structural Biotechnology Journal* **13**, 8–17.
URL: *https://www.sciencedirect.com/science/article/pii/S2001037014000464*

Ladhani, Shamez N., Georgina Ireland, Frances Baawuah, Joanne Beckmann, Ifeanyichukwu O. Okike, Shazaad Ahmad, Joanna Garstang, Andrew J. Brent, Bernadette Brent, Felicity Aiano, Zahin Amin-Chowdhury, Meaghan Kall, Ray Borrow, Ezra

Linley, Maria Zambon, John Poh, Lenesha Warrener, Angie Lackenby, Joanna Ellis, Gayatri Amirthalingam, Kevin E. Brown e Mary E. Ramsay ,(2021), 'Emergence of sars-cov-2 alpha (b.1.1.7) variant, infection rates, antibody seroconversion and seroprevalence rates in secondary school students and staff: Active prospective surveillance, december 2020 to march 2021, england', *Journal of Infection* **83**(5), 573–580.
URL: *https://doi.org/10.1016/j.jinf.2021.08.019*

Lamb, Kieran D., Joseph Hughes, Spyros Lytras, Francesca Young, Orges Koci, James Herzig, Simon C Lovell, Joe Grove, Ke Yuan e David L. Robertson ,(2024), 'From a single sequence to evolutionary trajectories: protein language models capture the evolutionary potential of sars-cov-2 protein sequences', *bioRxiv* .
URL: *https://www.biorxiv.org/content/early/2024/09/18/2024.07.05.602129*

Lauring, Adam S. e Emma B. Hodcroft ,(2021), 'Genetic variants of sars-cov-2—what do they mean?', *JAMA* **325**(6), 529–531.
URL: *https://doi.org/10.1001/jama.2020.27124*

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So e Jaewoo Kang ,(2019), 'Biobert: a pre-trained biomedical language representation model for biomedical text mining', *CoRR* **abs/1901.08746**.
URL: *http://arxiv.org/abs/1901.08746*

Li, Q., Z. Hu, Y. Wang, L. Li, Y. Fan, I. King, G. Jia, S. Wang, L. Song e Y. Li ,(2024), 'Progress and opportunities of foundation models in bioinformatics', *Brief Bioinform* **25**(6), bbae548.
URL: *https://academic.oup.com/bib/article/25/6/bbae548/7842778*

Liu, Y., J. Liu, K. S. Plante, J. A. Plante, X. Xie, X. Zhang, Z. Ku, Z. An, D. Scharton, C. Schindewolf, S. G. Widen, V. D. Menachery, P. Y. Shi e S. C. Weaver ,(2022), 'The n501y spike substitution enhances sars-cov-2 infection and transmission', *Nature* **602**(7896), 294–299.
URL: *https://doi.org/10.1038/s41586-021-04245-0*

Lloyd, S. P. ,(1982), 'Least squares quantization in pcm', *IEEE Transactions on Information Theory* **28**(2), 129–137.
URL: *https://doi.org/10.1109/TIT.1982.1056489*

MacQueen, J. B. ,(1967), Some methods for classification and analysis of multivariate observations, *em* 'Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability', Vol. 1, University of California Press, pp. 281–297.
URL: *https://digitalassets.lib.berkeley.edu/math/ucb/text/math_s5_v1_article-17.pdf*

McInnes, Leland, John Healy e James Melville ,(2020), 'Umap: Uniform manifold approximation and projection for dimension reduction'.
URL: *https://arxiv.org/abs/1802.03426*

Mutawa, A. M. ,(2025), 'Attention-based hybrid deep learning models for classifying covid-19 genome sequences', *AI* **6**(1), 4.
URL: *https://www.mdpi.com/2673-2688/6/1/4*

Nagpal, Sargun, Ridam Pal, Ashima, Ananya Tyagi, Sadhana Tripathi, Aditya Nagori, Saad Ahmad, Hara Prasad Mishra, Rishabh Malhotra, Rintu Kutum e Tavpritesh Sethi ,(2022), 'Genomic surveillance of covid-19 variants with language models and machine learning', *Frontiers in Genetics* **13**, 858252.
URL: *https://www.frontiersin.org/articles/10.3389/fgene.2022.858252/full*

Opitz, Juri ,(2024), 'A closer look at classification evaluation metrics and a critical reflection of common evaluation practice', *Transactions of the Association for Computational Linguistics* **12**, 820–836.
URL: *https://doi.org/10.1162/tacl_a_00675*

Pennington, Jeffrey, Richard Socher e Christopher Manning ,(2014), GloVe: Global vectors for word representation, *em* A.Moschitti, B.Pang e W.Daelemans, eds., 'Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, Doha, Qatar, pp. 1532–1543.
URL: *https://aclanthology.org/D14-1162/*

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee e Luke Zettlemoyer ,(2018), Deep contextualized word representations, *em* M.Walker, H.Ji e A.Stent, eds., 'Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)', Association for Computational Linguistics, New Orleans, Louisiana, pp. 2227–2237.
URL: *https://aclanthology.org/N18-1202/*

Pilehvar, Mohammad Taher e Nigel Collier ,(2016), De-conflated semantic representations, *em* J.Su, K.Duh e X.Carreras, eds., 'Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Austin, Texas, pp. 1680–1690.
URL: *https://aclanthology.org/D16-1174/*

Powers, David M. ,(2011), 'Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation', *Journal of Machine Learning Technologies* **2**(1), 37–63.
URL: *https://arxiv.org/abs/2010.16061*

Promja, S., J. Puenpa, T. Achakulvisut, Y. Poovorawan, S. Y. Lee, P. Athamanolap e B. Lertanantawong ,(2023), 'Machine learning-assisted real-time polymerase chain reaction and high-resolution melt analysis for sars-cov-2 variant identification', *Analytical Chemistry* **95**(3), 2102–2109.
URL: *https://doi.org/10.1021/acs.analchem.2c05112*

Qayyum, Abdul, Abdesslam Benzinou, Oumaima Saidani, Fatimah Alhayan, Muhammad Attique Khan, Anum Masood e Moona Mazher ,(2024), 'Assessment and classification of covid-19 dna sequence using pairwise features concatenation from multi-transformer and deep features with machine learning models', *SLAS Technology* **29**(4), 100147.
URL: *https://www.sciencedirect.com/science/article/pii/S2472630324000293*

Qin, J., X. Tian, S. Liu, Z. Yang, D. Shi, S. Xu e Y. Zhang ,(2024), 'Rapid classification of sars-cov-2 variant strains using machine learning-based label-free sers strategy', *Talanta* **267**, 125080.
URL: *https://pubmed.ncbi.nlm.nih.gov/37678002/*

Ranzani, Otavio T, Leonardo SL Bastos, Jose G M Gelli, Juliana F Marchesi, Fernanda Baião, Silvana Hamacher e Fernando A Bozza ,(2021), 'Characterisation of the first 250,000 hospital admissions for covid-19 in brazil: a retrospective analysis of nationwide data', *The Lancet Respiratory Medicine* **9**(4), 407–418. Epub 2021 Jan 15.
URL: *https://doi.org/10.1016/S2213-2600(20)30560-9*

Reimers, Nils e Iryna Gurevych ,(2019), 'Sentence-bert: Sentence embeddings using siamese bert-networks'.
URL: *https://arxiv.org/abs/1908.10084*

Rousseeuw, Peter J ,(1987), 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics* **20**, 53–65.
URL: *https://www.sciencedirect.com/science/article/pii/0377042787901257*

Scarpa, Fabio, Francesco Branda, Nicola Petrosillo e Massimo Ciccozzi ,(2024), 'On the sars-cov-2 variants', *Infectious Disease Reports* **16**(2), 289–297. Epub 2024 Mar 26.
URL: *https://doi.org/10.3390/idr16020024*

Singh, R., S. Nagpal, N. K. Pinna e S. S. Mande ,(2022), 'Tracking mutational semantics of sars-cov-2 genomes', *Scientific Reports* **12**(1), 15704.
URL: *https://doi.org/10.1038/s41598-022-20000-5*

Sokhansanj, Bahrad A., Zhengqiao Zhao e Gail L. Rosen ,(2022), 'Interpretable and predictive deep neural network modeling of the sars-cov-2 spike protein sequence to predict covid-19 disease severity', *Biology* **11**(12).
URL: *https://www.mdpi.com/2079-7737/11/12/1786*

Sokolova, Marina e Guy Lapalme ,(2009), 'A systematic analysis of performance measures for classification tasks', *Information Processing & Management* **45**(4), 427–437.
URL: *https://www.sciencedirect.com/science/article/pii/S0306457309000259*

Tegally, Houriiyah, Eduan Wilkinson, Marta Giovanetti, Arash Iranzadeh, Vagner Fonseca, Jennifer Giandhari, Deelan Doolabh, Sureshnee Pillay, James E San, Nokukhanya Msomi et al. ,(2021), 'Detection of a sars-cov-2 variant of concern in

south africa', *Nature* **592**(7854), 438–443.
URL: *https://www.nature.com/articles/s41586-021-03402-9*

Tervent, Juan R., Diana M. Cordova-Esparza, Alfonzo Ramirez-Pedraza e Edgar A. Chavez-Urbiola ,(2025), 'A comprehensive survey of loss functions and performance metrics in deep learning', *Artificial Intelligence Review* **58**.
URL: *https://link.springer.com/article/10.1007/s10462-025-11198-7*

van der Maaten, Laurens e Geoffrey Hinton ,(2008), 'Visualizing data using t-sne', *Journal of Machine Learning Research* **9**(86), 2579–2605.
URL: *http://jmlr.org/papers/v9/vandermaaten08a.html*

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser e Illia Polosukhin ,(2017), 'Attention is all you need', *CoRR* **abs/1706.03762**.
URL: *http://arxiv.org/abs/1706.03762*

Verma, Rohit Kumar, Rakesh Tiwari e Pratik Singh Thakur ,(2023), 'Partition coefficient and partition entropy in fuzzy c means clustering', *Journal of Scientific Research and Reports* **29**(12), 1–6.
URL: *https://doi.org/10.9734/jsrr/2023/v29i121812*

Villoutreix, Bruno O., Vincent Calvez, Anne-Geneviève Marcelin e Abdel-Majid Khatib ,(2021), 'In silico investigation of the new uk (b.1.1.7) and south african (501y.v2) sars-cov-2 variants with a focus at the ace2–spike rbd interface', *International Journal of Molecular Sciences* **22**(4).
URL: *https://www.mdpi.com/1422-0067/22/4/1695*

Wang, Jiao, Vivek Chavda, Riddhi Prajapati, Anjali Bedse, Jinita Patel, Sagar Popat, Gargi Jogi, Lakshmi Vineela Nalla, Keshava Jetha, Bairong Shen e Rajeev K. Singla ,(2023), 'An amalgamation of bioinformatics and artificial intelligence for covid-19 management: From discovery to clinic', *Current Research in Biotechnology* **6**, 100159.
URL: *https://www.sciencedirect.com/science/article/pii/S2590262823000412*

Wang, Pengfei, Manoj S. Nair, Lihong Liu, Sho Iketani, Yang Luo, Yicheng Guo, Maple Wang, Jian Yu, Baoshan Zhang, Peter D. Kwong, Barney S. Graham, John R. Mascola, Jennifer Y. Chang, Michael T. Yin, Magdalena Sobieszczyk, Christos A. Kyratsous, Lawrence Shapiro, Zizhang Sheng, Yaoxing Huang e David D. Ho ,(2021), 'Antibody resistance of sars-cov-2 variants b.1.351 and b.1.1.7', *Nature* **593**(7857), 130–135.
URL: *https://doi.org/10.1038/s41586-021-03398-2*

Wang, Wenhui, Furu Wei, Li Dong, Hangbo Bao, Nan Yang e Ming Zhou ,(2020), 'Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers', *arXiv preprint arXiv:2002.10957* .
URL: *https://arxiv.org/abs/2002.10957*

Webster, Jonathan J. e Chunyu Kit ,(1992), 'Tokenization as the initial phase in nlp', *Proceedings of the 14th conference on Computational linguistics - Volume 4* pp. 1106–1110.
URL: *https://doi.org/10.3115/992424.992434*

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest e Alexander Rush ,(2020), Transformers: State-of-the-art natural language processing, *em* Q.Liu e D.Schlangen, eds., 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations', Association for Computational Linguistics, Online, pp. 38–45.
URL: *https://aclanthology.org/2020.emnlp-demos.6/*

World Health Organization ,(2025), 'Who coronavirus (covid-19) dashboard', `https://covid19.who.int/`. Accessed in April 2025. Available at: `https://covid19.who.int/`.

Xie, X.L. e G. Beni ,(1991), 'A validity measure for fuzzy clustering', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**(8), 841–847.
URL: *https://doi.org/10.1109/34.85677*

Ye, Qiushi, He Wang, Fanding Xu, Sijia Zhang, Shengli Zhang, Zhiwei Yang e Lei Zhang ,(2024), 'Co-mutations and possible variation tendency of the spike rbd and membrane protein in sars-cov-2 by machine learning', *International Journal of Molecular Sciences* **25**(9).
URL: *https://www.mdpi.com/1422-0067/25/9/4662*

Zhou, Binbin, Hang Zhou, Xue Zhang, Xiaobin Xu, Yi Chai, Zengwei Zheng, Alex Chichung Kot e Zhan Zhou ,(2023), 'Tempo: A transformer-based mutation prediction framework for sars-cov-2 evolution', *Computers in Biology and Medicine* **152**, 106264.
URL: *https://www.sciencedirect.com/science/article/pii/S0010482522009726*

Zvyagin, Maxim, Alexander Brace, Kyle Hippe, Yuntian Deng, Bin Zhang, Cindy Orozco Bohorquez, Austin Clyde, Bharat Kale, Danilo Perez-Rivera, Heng Ma, Carla M. Mann, Michael Irvin, J. Gregory Pauloski, Logan Ward, Valerie Hayot-Sasson, Murali Emani, Sam Foreman, Zhen Xie, Diangen Lin, Maulik Shukla, Weili Nie, Josh Romero, Christian Dallago, Arash Vahdat, Chaowei Xiao, Thomas Gibbs, Ian Foster, James J. Davis, Michael E. Papka, Thomas Brettin, Rick Stevens, Anima Anandkumar, Venkatram Vishwanath e Arvind Ramanathan ,(2022), 'Genslms: Genome-scale language models reveal sars-cov-2 evolutionary dynamics', *bioRxiv* .
URL: *https://www.biorxiv.org/content/early/2022/11/23/2022.10.10.511571*