Previsão baseada em aprendizado de máquina do diagnóstico de SRAG, incluindo COVID-19, influenza e outros, com base nos sintomas

Pamela Rocha[1], **Luiz Ferreira**[2], **Victória**[3]

[1]Analise e Desenvolvimento de Sistemas
Faculdade NovaRoma,
Caruaru – PE – Brasil

Abstract. *Among severe respiratory diseases, Severe Acute Respiratory Syndrome (SARS) stands out for its severity and diagnostic complexity, especially in pandemic outbreaks such as Influenza and COVID-19. This article explores the performance of applying artificial intelligence (AI) and machine learning (ML) techniques for the pre-diagnosis of influenza, COVID-19, or others, using epidemiological data provided by the Ministry of Health. Methodological steps include data preprocessing, class balancing, feature selection, ML model development, training, and testing using accuracy metrics and Cross-Validation to confirm its effectiveness. The use of algorithms such as Random Forest has shown potential to assist in screening and rapid diagnosis, improving the health system's response and contributing to the effectiveness of epidemiological surveillance.*

Resumo. *Dentre as doenças respiratórias graves, a Síndrome Respiratória Aguda Grave (SRAG) destaca-se pela sua gravidade e complexidade de diagnóstico, especialmente em contextos de surtos pandêmicos como os da Influenza e COVID-19. Este artigo explora o desempenho da aplicação de inteligência artificial (IA) e técnicas de aprendizado de máquina (ML) para o pré-diagnóstico de influenza, COVID-19 ou outros, utilizando dados epidemiológicos fornecidos pelo Ministério da Saúde. As etapas metodológicas incluem o pré-processamento dos dados, balanceamento das classes, seleção de características, desenvolvimento de modelos de ML, treinamento e testes usando métricas de acurácia e de Validação Cruzada para confirmar sua eficácia. O uso de algoritmos como Random Forest demonstrou potencial para auxiliar na triagem e diagnóstico rápido, melhorando a resposta do sistema de saúde e contribuindo para a eficácia da vigilância epidemiológica.*

1. Introdução

A SRAG (Síndrome Respiratória Aguda Grave) é uma condição médica grave que consiste no prejuízo do sistema respiratório, caracterizada por sintomas agudos no trato respiratório, dispneia (falta de ar), desconforto respiratório, pressão ou dor persistente na região do tórax(Secretaria de Estado da Saúde de Santa Catarina, s.d.)

Diante disto, este artigo propõe-se a explorar os bancos de dados epidemiológicos da SRAG, disponibilizados pelo Ministério da Saúde, através do portal oficial do governo federal brasileiro. Mas a princípio, é importante compreendermos do que se trata a influenza e a COVID-19.

Primeiramente vamos entender que na escala geográfica existem diferenças no que se refere a pandemia, epidemia ou surto epidemiológico, a suas diferenças estão no número de pessoas afetadas: surtos são locais, epidemias são regionais e pandemias são globais.

Seguindo a ordem cronológica, 2009 ficou marcado pelo início da pandemia de influenza, que foi popularmente chamada de gripe suína, pois seu surgimento se deu pela contaminação através dos porcos, que já sofriam dessa infecção respiratória. De acordo com falas do médico brasileiro Drauzio Varella em 2011(Varella, Drauzio, s.d.) "A gripe H1N1, ou influenza A, é provocada pelo vírus H1N1, um subtipo da influenza vírus do tipo A. Ele é resultado da combinação de segmentos genéticos do vírus humano da gripe, do vírus da gripe aviária e do vírus da gripe suína".

Os sinais e sintomas da influenza podem incluir febre alta, tosse, dores de garganta, dores musculares, fadiga, dores de cabeça, calafrios, congestão nasal, cansaço extremo e, às vezes, vômitos e diarreia. A transmissão entre seres humanos ocorre de forma aeróbia, através das gotículas salivares, resultantes de tosses, espirros ou falas. De acordo com as autoras Bellei e Melchior, em 2011,(VARGAS et al., 2022a) "No Brasil, 2.051 óbitos e mais de 44 mil casos da doença foram confirmados no mesmo ano". As autoras também complementam que, de acordo com a OMS, a pandemia de influenza foi finalizada em 10 de agosto de 2010. De maneira mais clara, a gripe A que diz respeito ao vírus H1N1, diferencia-se da COVID-19, que não pertence à mesma categoria taxonômica, ou seja, não é a mesma doença nem o mesmo vírus.

Foi no final do ano de 2019, que se iniciou a pandemia da COVID-19 (Coronavirus Disease 2019), causada pelo agente SARS-CoV-2, que pertence a uma ampla "família" de vírus distinta da influenza. Quando um indivíduo é infectado, os primeiros sinais e sintomas apresentados assemelham-se a um resfriado comum; entretanto, a COVID-19 pode agir de forma diferente da gripe. Os sintomas leves podem consistir em febre, tosse, fadiga, dor de garganta, dor de cabeça, perda de olfato ou paladar e congestão nasal. Já os sintomas agravados incluem dificuldade intensa para respirar ou desconforto respiratório, pressão duradoura no peito, saturação de oxigênio no sangue inferior a 95% e coloração azulada ou acinzentada da pele, decorrente de oxigenação insuficiente do sangue (cianose) segundo dados da Secretaria de Estado de Saúde de Minas Gerais, em 2020. Já o número de óbitos registrados pela doença, atualmente registra mais de 712 mil mortes no Brasil.(Dados.gov.br, s.d.)

Ambas as pandemias resultaram em elevados números de casos de contaminações e óbitos. Com a alta transmissibilidade e a severidade dessas doenças, o sistema de saúde

teve a necessidade de aumento na assistência à saúde pública, tais como; hospitais de campanha e aumento da verba destinada a estudos para o desenvolvimento de vacinas, com o objetivo de controlar a contaminação e evolução dos casos, os quais geraram uma sobrecarga no sistema de saúde. Atualmente, o MS tem investido na vacinação da população em geral, mantendo assim o controle sobre as contaminações da influenza e COVID – 19. Porém, ainda existem registros de novos casos das doenças e consequentemente das SRAG, o que destaca a necessidade de um diagnóstico rápido e preciso para otimizar os recursos e melhorar o controle dos vírus agentes, ainda atualizados na base do governo, já recebendo atualização do ano de 2024.

Durante a última pandemia, cidades adotaram a tecnologia para teleconsultas, que consiste no atendimento médico remoto, a fim de evitar essas sobrecargas. A faculdade de medicina UFMG, em 2020, destaca que "o uso da tecnologia se tornou uma alternativa ao atendimento presencial para diminuir o risco de contaminação pelo coronavírus dentro das unidades de saúde". Analisando as situações deste último ano (2023), houve uma explosão no uso da inteligência artificial (IA) (Insper, s.d.)

A integração da IA no diagnóstico de SRAG pode representar um avanço significativo para o sistema de saúde brasileiro, que pode auxiliar em responder às emergências de saúde pública. Por tanto, este estudo se propõe a desenvolver e implementar as IA's para serem capazes de dar uma estatística do possível diagnóstico de influenza, covid-19 ou outros, utilizando as técnicas de aprendizagem de máquina (ML). Este projeto de IA e BIGDATA será treinado para ler o banco de dados fornecido pelo MS e interpretar em estatística o possível diagnóstico, com o intuito de auxiliar o profissional de saúde para o pre-diagnóstico, dando ao paciente uma maior rapidez para tratamento, assim evitando que o caso evolua para a SRAG. A iniciativa visa melhorar a vigilância epidemiológica e o controle de doenças graves como a influenza e a COVID-19, contribuindo para a eficiência do sistema de saúde. Para alcançar este objetivo, o estudo envolverá a exploração de bancos de dados epidemiológicos disponibilizados pelo Ministério da Saúde, além de múltiplas etapas que incluem a coleta e processamento de dados, desenvolvimento, treinamento e avaliação do modelo de IA.

2. Fundamentação Teórica

2.1. Tema Principal

Este trabalho se baseia em dados da Secretaria de Vigilância em Saúde (SVS), que desenvolveu e implantou a vigilância da SRAG no Brasil, com o objetivo de monitorar as pandemias de 2009 e 2019.

O estudo analisa os dados de vigilância da SRAG no Brasil, utilizando informações fornecidas pela SVS, que monitora as pandemias de H1N1 e COVID-19. As triagens médicas são disponibilizadas no Sistema de Informações de Agravos de Notificação (SINAN), no site do governo federal.

O artigo propõe utilizar os dados epidemiológicos do Ministério da Saúde para desenvolver e implementar um sistema de IA capaz de interpretar estatisticamente os possíveis diagnósticos. Este guia detalhado descreve, passo a passo, a metodologia para pré-processar, equilibrar, selecionar características e criar modelos de ML utilizando dados do SINAN.

2.2. Metricas e Técnicas

1. **Pré-Processamento:** O pré-processamento de uma base de dados é uma etapa crucial para garantir a qualidade dos dados e, consequentemente, a eficácia das análises subsequentes. Esse processo envolve a limpeza dos dados, como remoção de valores ausentes ou inconsistentes, e a transformação, que pode incluir normalização e codificação de variáveis. Além disso, o pré-processamento facilita a detecção de padrões e a melhoria do desempenho dos algoritmos de aprendizado de máquina. Em resumo, realizar o pré-processamento de dados é essencial para obter resultados mais precisos e confiáveis nas análises e modelos preditivos.(BATISTA et al., 2003).

   Neste trabalho foram adotados a limpeza da base de dados baseado em sua quantidade de valores nulos por linhas e features, estas foram contabilizadas e descartadas quando o total de valores nulos chegaram em 80% dos valores totais contados, resumindo assim as informações da base de dados.

2. **Balanceamento dos Dados:** A princípio realizar o balanceamento de dados é uma técnica fundamental em machine learning para lidar com conjuntos de dados onde as classes que queremos prever têm quantidades desproporcionais, o que prejudicaria os resultados tornando o modelo enviesado. Ajustar a distribuição das classes no conjunto de dados para garantir que o modelo de aprendizado de máquina não seja tendencioso em relação a uma classe dominante. Após a limpeza dos dados, onde foi usada as técnicas descritas abaixo, foram geradas 2 (duas) bases de dados distintas e balanceadas, que serão utilizadas para treinar e testar os modelos escolhidos para o projeto.(SANTIAGO, 2021)

   2.1 **SMOTE:** É o algoritmo de pré-processamento da Técnica de Sobreamostragem Minoritária Sintética (SMOTE) é utilizado para aumentar o número de casos em um conjunto de dados de forma equilibrada, corrigindo o desbalanceamento das classes. Dados desequilibrados (ou desbalanceados) ocorrem quando as classes no conjunto de dados de treinamento não estão igualmente representadas. Isso significa que algumas classes têm mais exemplos do que outras. Primeiramente, é necessário identificar a classe minoritária e a quantidade de suas instâncias. Em seguida, a técnica de reamostragem a classe majoritária para equilibrar o número de instâncias entre as classes, evitando que o modelo aprenda padrões enviesados.(WANG et al., 2006)

   2.2 **Undersampling:** A técnica de undersampling é utilizada para equilibrar conjuntos de dados desbalanceados, onde uma classe é significativamente mais representada do que outra. Isso é feito reduzindo o número de exemplos da classe majoritária, tornando o conjunto de dados mais equilibrado. Apesar de ajudar a melhorar a performance de algoritmos de aprendizado de máquina ao evitar vieses, essa técnica pode levar à perda de informações importantes se não for aplicada com cuidado. Portanto, é crucial usar o undersampling de maneira estratégica para

manter a integridade dos dados e garantir resultados mais justos e precisos.(MOHAMMED; RAWASHDEH; ABDULLAH, 2020)

3. **Feature Selection:** A seleção de recursos é uma técnica que escolhe os melhores subconjunto de características, removendo recursos ruidosos, irrelevantes e redundantes, assim reduzindo a dimensionalidade do conjunto de dados. Este modelo é construído com a Seleção Sequencial de Atributos (SFS) e a Seleção Sequencial Reversa (SBS), para identificar quais características são mais importantes para o modelo utilizado. A Seleção Sequencial de Atributos (SFS) auxilia o modelo a selecionar características, reduzindo a dimensionalidade e potencialmente aumentando a precisão. Já a Seleção Sequencial Reversa Sequential Backward Selection (SBS) começa com todas as características do modelo e remove as menos importantes, até encontrar o subconjunto ótimo que maximiza a acurácia do modelo.(VALANDRO, 2021)

4. **Random Forest (Decision Trees):** Random Forest trata-se de um método de ensemble learning que combina múltiplas árvores de decisão para melhorar a precisão e reduzir o overfitting, que ocorre quando o modelo tem um desempenho excelente nos dados de treinamento, mas seu desempenho é ruim em dados novos e não vistos (dados de teste ou validação). Utilizado como modelo de classificação após o balanceamento dos dados e a seleção de características. Ele é composto por uma ”floresta”de árvores de decisão, onde cada árvore é treinada com um subconjunto diferente do conjunto de dados e com um subconjunto diferente de características.(EBAC Online, 2021)

5. **Naive Bayes:** É um método para tarefas de classificação, especialmente quando as características são independentes ou quase independentes. No projeto, o Naive Bayes é usado para treinar um modelo de classificação com os dados de SRAG, demonstrando como ele pode ser aplicado em um problema real. O Teorema de Bayes descreve a probabilidade de um evento com base no conhecimento prévio de condições relacionadas ao evento (wikipedia, 2023).(Wikipedia, 2021)

6. **K-NN:** K-nearest neighbors algorithm, é um método de aprendizagem supervisionado não paramétrico, utilizado para classificação e regressão. Ele simplesmente armazena o conjunto de dados de treinamento, que é composto por pontos de dados rotulados. Na classificação ele identifica todas as distâncias calculadas e seleciona os “K” pontos de dados do conjunto de treinamento que estão mais próximos da nova amostra, já na regressão ao invés de votar nos valores dos “K” vizinhos mais próximos são simplesmente promediados para dar a previsão final.(DataGeeks, 2021)

2.2.1. Trabalhos relacionados:

Esta seção apresenta uma revisão dos principais trabalhos relacionados ao uso de algoritmos de aprendizado de máquina na previsão de desfechos clínicos relevantes para a saúde pública. Inicialmente, discutiremos estudos que abordam a predição de síndromes e complicações médicas, seguidos por pesquisas que investigam a aplicação desses modelos na triagem e diagnóstico de doenças.

1. Machine learning for the early prediction of acute respiratory distress syndrome (ARDS) in patients with sepsis in the ICU based on clinical data.(YIN et al., 2024).

2. Supervised machine learning for the early prediction of acute respiratory distress syndrome (ARDS)(D'AMATO et al., 2020).

2.2.2. Predição de Síndromes e Complicações Médicas

Um estudo conduzido em. (2024), propôs um modelo de predição precoce para a síndrome do desconforto respiratório agudo (SDRA) em pacientes com sepse. Os pesquisadores utilizaram algoritmos de aprendizado de máquina e dados clínicos para desenvolver um modelo com alta precisão na identificação de pacientes em risco de desenvolver SDRA. Em um estudo subsequente conduzido em (2020) e publicado no Journal of Forensic and Legal Medicine, pesquisadores utilizaram uma abordagem similar para desenvolver um modelo preditivo para a síndrome do desconforto respiratório agudo (SDRA) em pacientes, demonstrando alta precisão na identificação precoce de pacientes em risco.(VARGAS et al., 2022b)

2.2.3. Aplicação na Triagem e Diagnóstico de Doenças

1. Effects of molecular weight on the antibacterial activity and mechanism of action of chitosan against Gram-negative bacteria(LI et al., 2021)

2. A novel AI-enabled framework to diagnose Coronavirus COVID-19 using smartphone embedded sensors: design study(MAGHDID et al., 2020)

Em um estudo recente, Yazeed Zoabi. (2021) investigou a eficácia de modelos de aprendizado de máquina na triagem de COVID-19 com base em características clínicas e histórico de exposição. Publicado na revista na Digital Medicine, o estudo desenvolveu um modelo que utilizou variáveis como idade, sexo, sintomas clínicos (por exemplo, febre, tosse), e histórico de contato com indivíduos infectados. O modelo alcançou uma alta precisão. Esses resultados são consistentes com os achados de Wei Jiang. (2021), que exploraram a aplicação de algoritmos de aprendizado de máquina na identificação de casos suspeitos de COVID-19 utilizando dados de tomografia computadorizada (TC) combinados com informações clínicas. Publicado na Scientific Reports, o estudo demonstrou

que a integração de dados clínicos e imagens médicas pode aumentar significativamente a precisão do diagnóstico.

2.2.4. Identificação de Lacunas

Apesar dos avanços significativos nos estudos anteriores, ainda existem lacunas a serem preenchidas na literatura. Poucos estudos investigaram a aplicação de modelos de aprendizado de máquina na predição de Identificação para SARS/SRAG. Este estudo visa preencher essa lacuna ao focar especificamente nesse grupo e examinar a eficácia dos modelos de aprendizado de máquina na predição de Identificação.

2.2.5. Conclusão da Seção

Em resumo, os estudos revisados destacam os avanços promissores no uso de algoritmos de aprendizado de máquina na previsão de desfechos clínicos e no auxílio ao diagnóstico médico, especialmente na identificação precoce de síndromes e complicações médicas, bem como na triagem e diagnóstico de doenças, como o COVID-19. No entanto, à medida que avançamos nessa área, é fundamental continuar explorando novos modelos e variáveis para melhorar a precisão e generalização dos resultados, visando assim contribuir cada vez mais para a saúde pública e o bem-estar dos pacientes.

3. Metodologia:

A metodologia adotada neste projeto segue os parâmetros da metodologia CRISP-DM, a qual predetermina uma forma analítica de processar e explorar dados por meio de análises. Composta por 6 fases distintas, estas geram um guia para o profissional/cientista de dados realizar a sua principal atribuição, que é resolver problemas, mantendo o foco. Contudo, o modelo permite uma reavaliação das etapas adotadas, tornando-se um modelo interativo para que estas etapas possam ser aprimoradas.(CHAPMAN et al., 1999)

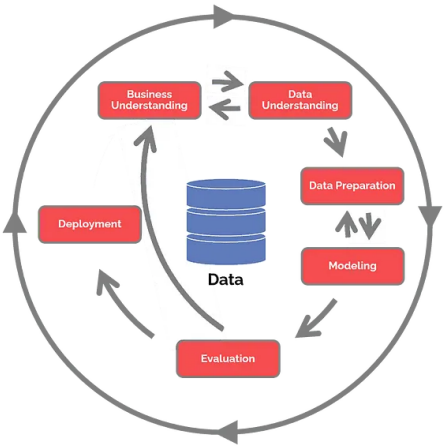

Figura 1. CRISP-DM

3.1. 1º Etapa - Compreensão do Negócio (Business Understanding)

Nesta etapa, vamos procurar compreender o contexto e os objetivos que o projeto deseja alcançar, buscando esclarecer os problemas e os critérios adotados para o sucesso do projeto. O processo de definição de um diagnóstico é árduo e de suma importância para o bem-estar e a melhora do paciente, sendo necessário que seja atribuído o mais cedo possível no ciclo de atendimento. Com isso em mente, iremos treinar um modelo com os dados de notificação de SRAG para auxiliar e acelerar o processo de diagnóstico. Utilizaremos diversas técnicas, que compreendem várias etapas, desde o pré-processamento e codificação até o treinamento de modelos de Machine Learning, baseados em bases de dados balanceadas usando técnicas específicas. Isso nos retornará modelos que, após treinados, poderão fornecer uma estimativa do possível diagnóstico.

3.2. 2º Etapa - Compreensão dos Dados (Data Understanding)

Nesta etapa, procuramos abordar os dados utilizados, bem como algumas possíveis lacunas ou potenciais problemas contidos neles. Este trabalho tem como base principal dados da Secretaria de Vigilância em Saúde (SVS). A base fornecida pelo SINAN consiste em perguntas e informações coletadas dos pacientes, com o objetivo de monitorar as pandemias de influenza em 2009 e a da COVID-19 a partir de 2020.(Ministério da Saúde, 2022).

Os dados utilizados incluem um Dicionário de Dados, que fornece mais informações sobre todas as colunas e descrições dos dados utilizados. O Dicionário de Dados pode ser acessado aqui. Entre as diversas colunas, destaca-se a coluna ”CLASSI_FIN”, que descreve a classificação final do caso e se torna nossa Coluna Alvo.

3.3. 3º Etapa - Preparação dos Dados (Data Preparation)

Esta etapa envolve o tratamento e a modelagem dos dados brutos utilizados no projeto. Aqui, realizamos um processo de limpeza dos valores nulos, contando os valores nulos de colunas e linhas e calculando sua porcentagem em relação ao total de colunas. Após calcular a porcentagem de valores nulos, as colunas e linhas que apresentaram mais de 80% de valores nulos foram removidas.

Após a remoção dos valores nulos, foi realizada uma verificação das features restantes utilizando o Dicionário de Dados fornecido com a base. A partir dessa verificação, observou-se que algumas colunas continham informações que não se alinhavam com o objetivo do projeto, como Região da Coleta, Nome do Hospital, Região do Paciente, entre outras informações regionais. Em seguida, criamos uma nova feature chamada “DIAS_SINTOMA” para a base fornecida. Esta feature foi criada calculando a diferença entre duas datas: a data da primeira notação dos sintomas e a data de coleta da informação. Isso representa o número de dias que o paciente está com aqueles sintomas que foram declarados na notificação.

Também codificamos valores binários, como a coluna “CS_SEXO”, que contém informações sobre o sexo do paciente. Outra abordagem adotada foi preencher os campos nulos ou vazios com um número distinto, o que se faz necessário para auxiliar os modelos que serão usados posteriormente.

Para finalizar, realizamos um processo de balanceamento dos dados, visto que, após a contagem das classes da coluna alvo “CLASSI_FIN”, observamos que estavam

desbalanceadas, causando overfitting nos treinamentos dos modelos adotados. Nesta etapa de balanceamento, adotamos duas técnicas distintas:

1. **SMOTE (Synthetic Minority Over-sampling Technique):** é uma técnica usada para lidar com conjuntos de dados desequilibrados em machine learning. Ele funciona gerando artificialmente novos exemplos da classe minoritária para equilibrar o número de exemplos entre as classes. O SMOTE cria novos exemplos interpolando entre os exemplos existentes da classe minoritária, selecionando aleatoriamente dois ou mais pontos próximos e gerando novos pontos ao longo da linha que os conecta. Isso ajuda a melhorar o desempenho dos modelos de machine learning, evitando que eles sejam tendenciosos em favor da classe majoritária.

2. **Undersampling:** é uma técnica usada para lidar com conjuntos de dados desequilibrados em machine learning. Ele funciona reduzindo o número de exemplos da classe majoritária para equilibrar o número de exemplos entre as classes. Isso é feito removendo aleatoriamente uma parte dos exemplos da classe majoritária até que as duas classes tenham tamanhos comparáveis. Embora o undersampling possa ajudar a melhorar o desempenho dos modelos, ele também pode levar à perda de informações importantes se muitos exemplos forem removidos.

Com isto foram criadas duas bases distintas, e estas são utilizadas alternadamente para treinamento e teste dos modelos adotados no trabalho.

3.4. 4º Etapa - Modelagem dos Dados (Modeling)

Esta etapa consiste em construir e treinar os modelos adotados, definindo suas abordagens e características particulares para o problema específico, buscando ajustá-los para garantir os melhores escores. As bases de dados foram consumidas diretamente do portal da Secretaria de Saúde(SRAG. . . , 2022).

Esses dados foram processados usando a ferramenta Google Colab, com a linguagem Python, utilizando bibliotecas como Pandas, Matplotlib, Classification Report, Confusion Matrix, entre outras. No início do projeto, foi definido que o problema abordado se enquadra nos moldes de classificação.

Adotamos três modelos para os treinamentos: Random Forest, Naive Bayes e K-NN. Esses modelos foram criados em pares, considerando que obtivemos duas bases distintas após o processo de balanceamento mencionado anteriormente. Assim, foram realizados dois treinamentos distintos para cada base, resultando em seis modelos diferentes. Após isso, esses modelos foram utilizados para realizar um processo de seleção de features (Feature Selection), utilizando a técnica SFS (Sequential Feature Selection), onde destas foram obtidos as melhores features que levaram ao maior índice de acurácia entre os modelos utilizados.

3.5. 5º Etapa - Avaliação do Modelo

Esta etapa se destaca por ser aquela em que descrevemos os resultados obtidos dos modelos treinados, destacando técnicas de Validação Cruzada e as Métricas de Desempenho.

A métrica de acurácia é preferida em problemas de classificação devido à sua simplicidade e facilidade de interpretação. Ela representa a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões, oferecendo uma visão geral do desempenho do modelo. No entanto, é importante estar ciente de que a acurácia pode ser enganosa em casos de desequilíbrio de classe, onde outras métricas, como precisão, recall ou F1-score, podem ser mais informativas. Em problemas com classes balanceadas, como será a nossa situação abordada, e de igual importância entre elas, a acurácia continua sendo uma métrica valiosa para avaliar o desempenho do modelo de classificação.

No projeto, realizamos a divisão da base de dados balanceada entre conjuntos de teste e treinamento, os quais são usados respectivamente pelos modelos para realização de Validação Cruzada entre as bases. Dados os modelos utilizados e a natureza do objetivo a ser abordado é a métrica escolhida foi a de Acurácia dos Modelos. Abaixo estão os resultados obtidos:

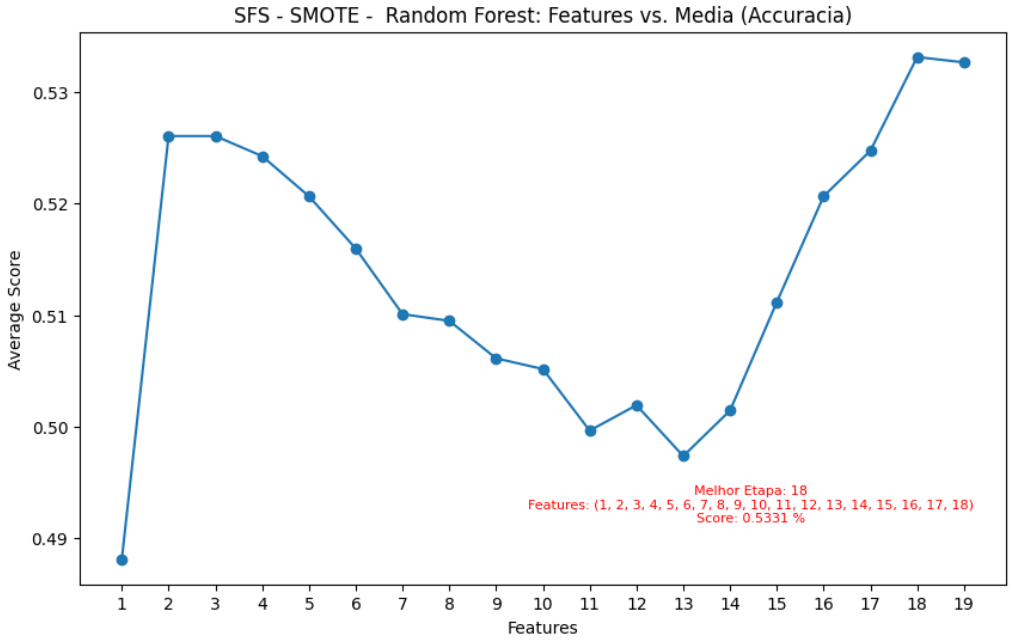

Figura 2. Modelo, SFS - Random Forest - SMOTE

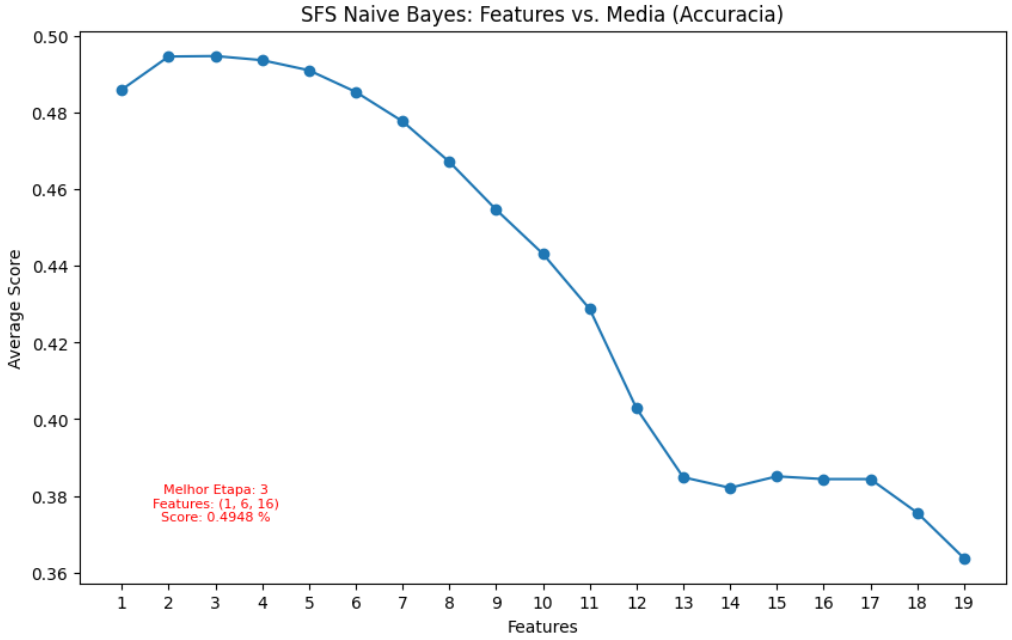

Figura 3. Modelo, SFS - Naive Bayes - SMOTE

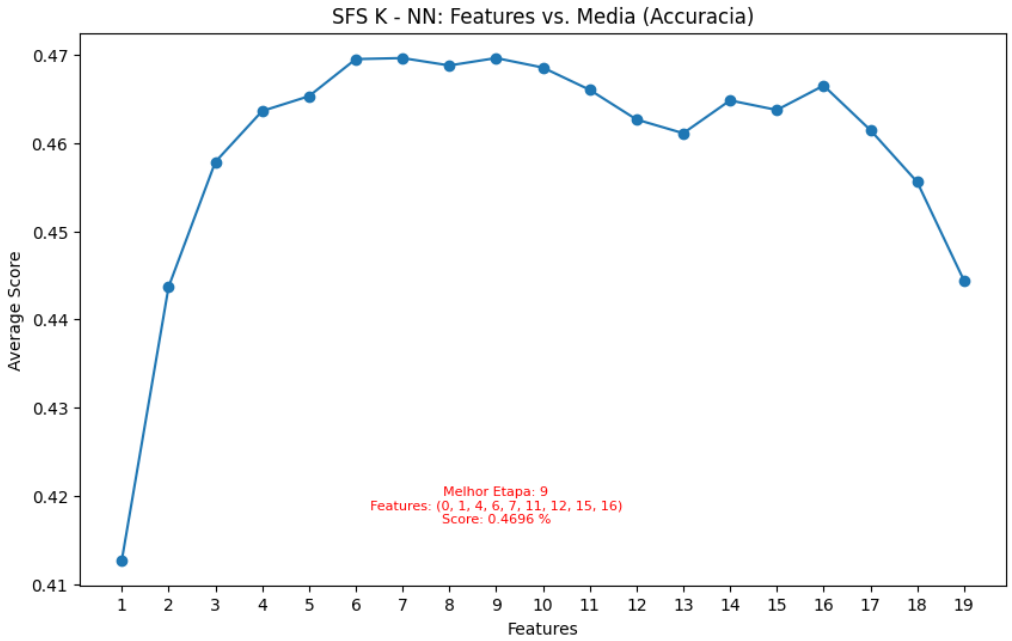

Figura 4. Modelo, SFS - Naive Bayes - SMOTE

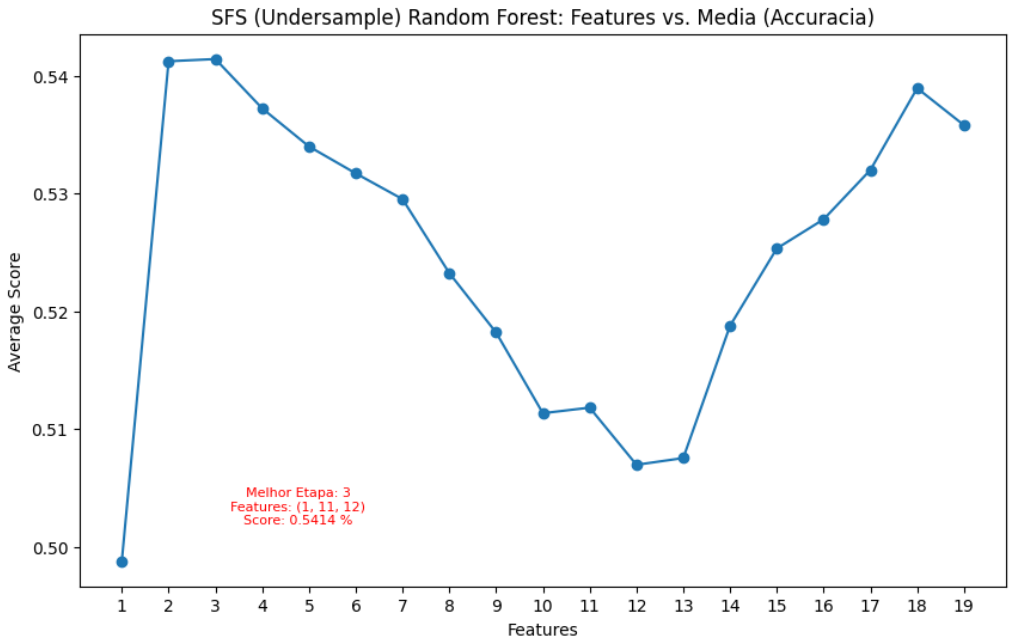

Figura 5. Modelo, SFS - Naive Bayes - SMOTE

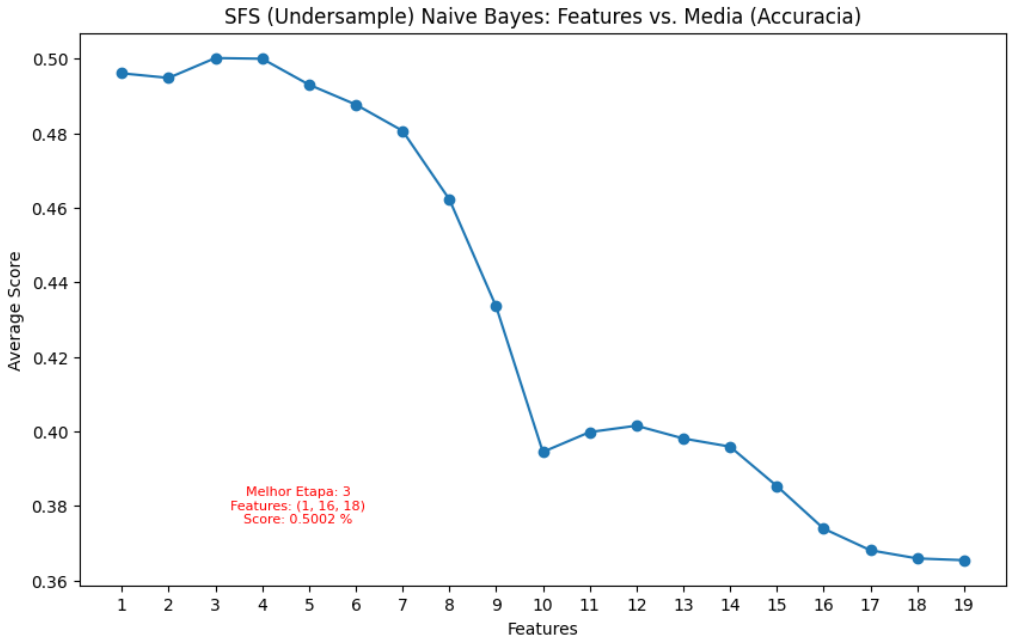

Figura 6. Modelo, SFS - Naive Bayes - SMOTE

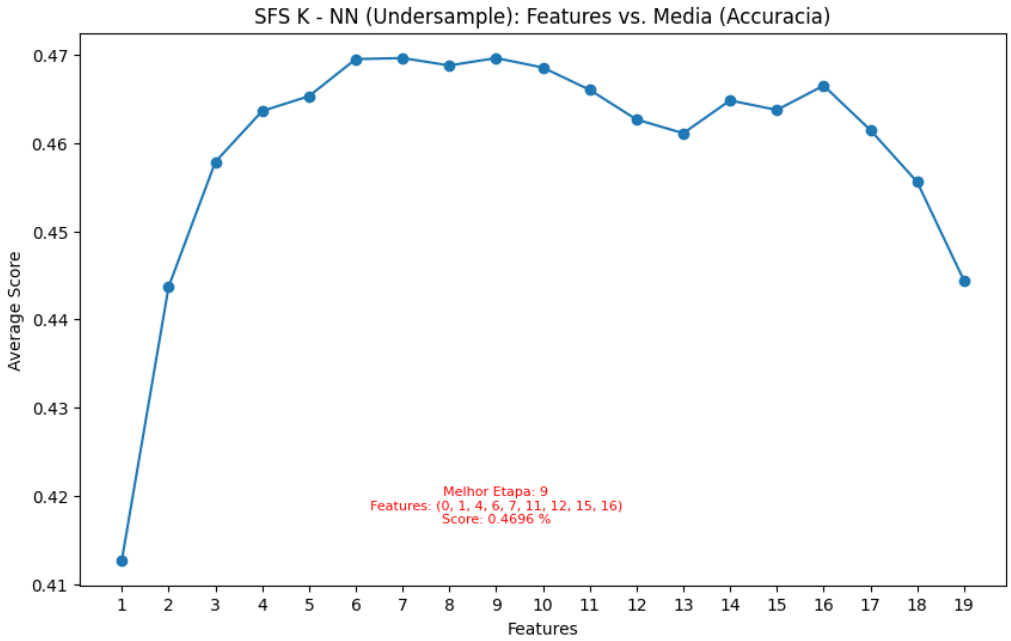

Figura 7. Modelo, SFS - Naive Bayes - SMOTE

Com base nas métricas coletadas durante as Validações Cruzadas e no Feature Selection, conseguimos obter informações valiosas que podemos utilizar para avaliar os modelos treinados. A partir dessas informações, selecionamos três modelos para passarem pelo processo de Random Search. Essa etapa visa aprimorar os hiperparâmetros utilizados nesses modelos, visando melhorar ainda mais o desempenho e a precisão das previsões.

Os modelos selecionados com base em suas respectivas acurácias foram o 'SFS - RandomSearch - SMOTE', o 'SFS - RandomSeach - Undersampling' e o 'SFS - Naive Bayes - Undersampling'. Esses modelos foram escolhidos devido às suas métricas de acurácia, que se destacaram entre os seis modelos testados. Após ajustes de seus hiperparâmetros, esses modelos serão testados, gerando os resultados descritos abaixo.

4. Resultados

Nesta seção, apresentaremos os resultados obtidos a partir da execução dos modelos descritos na seção anterior. Serão discutidos os cenários nos quais os modelos foram testados, os resultados da otimização, as métricas de desempenho, a construção do protótipo, as tecnologias utilizadas, bem como informações sobre o repositório do projeto e links para acesso.

Repositorio GitHub (SRAG, 2024)

Classe	**Precisão**	**Recall**	**F1-Score**	**Support**
SRAG por influenza	0.54	0.50	0.52	695
SRAG por outro vírus respiratório	0.53	0.50	0.51	713
SRAG por COVID-19	0.52	0.60	0.55	674
Acurácia			0.53	2082
Média	0.53	0.53	0.53	2082
Ponderada			0.53	2082

Tabela 1. Cenário 1: SFS – Random Forest - Treinado com SMOTE.

Classe	**Precisão**	**Recall**	**F1-Score**	**Support**
SRAG por influenza	0.40	0.03	0.06	695
SRAG por outro vírus respiratório	0.36	0.93	0.52	713
SRAG por COVID-19	0.41	0.10	0.16	674
Acurácia			0.36	2082
Média	0.39	0.36	0.25	2082
Ponderada			0.25	2082

Tabela 2. Cenário 2: SFS - Naive Bayes - Treinado com SMOTE.

Classe	Precisão	Recall	F1-Score	Support
SRAG por influenza	0.39	0.45	0.42	695
SRAG por outro vírus respiratório	0.47	0.48	0.47	713
SRAG por COVID-19	0.47	0.38	0.42	674
Acurácia			0.44	2082
Média	0.44	0.44	0.44	2082
Ponderada			0.44	2082

Tabela 3. Cenário 3: SFS - K NN - Treinado com SMOTE.

Classe	Precisão	Recall	F1-Score	Support
SRAG por influenza	0.52	0.38	0.44	869
SRAG por outro vírus respiratório	0.53	0.49	0.51	858
SRAG por COVID-19	0.53	0.71	0.61	889
Acurácia			0.53	2626
Média	0.53	0.53	0.52	2626
Ponderada			0.53	2626

Tabela 4. Cenário 4: SFS - Random Forest - Treinado com Undersampling

Classe	Precisão	Recall	F1-Score	Support
SRAG por influenza	0.48	0.24	0.32	869
SRAG por outro vírus respiratório	0.53	0.55	0.54	858
SRAG por COVID-19	0.51	0.74	0.60	889
Acurácia			0.51	2626
Média	0.50	0.51	0.49	2626
Ponderada			0.50	2626

Tabela 5. Cenário 5: SFS - Naive Bayes - Treinado com Undersampling.

Classe	Precisão	Recall	F1-Score	Support
SRAG por influenza	0.43	0.51	0.46	869
SRAG por outro vírus respiratório	0.42	0.38	0.40	858
SRAG por COVID-19	0.53	0.48	0.50	889
Acurácia			0.46	2626
Média	0.46	0.46	0.45	2626
Ponderada			0.46	2626

Tabela 6. Cenário 6: SFS - K NN - Treinado com Undersampling

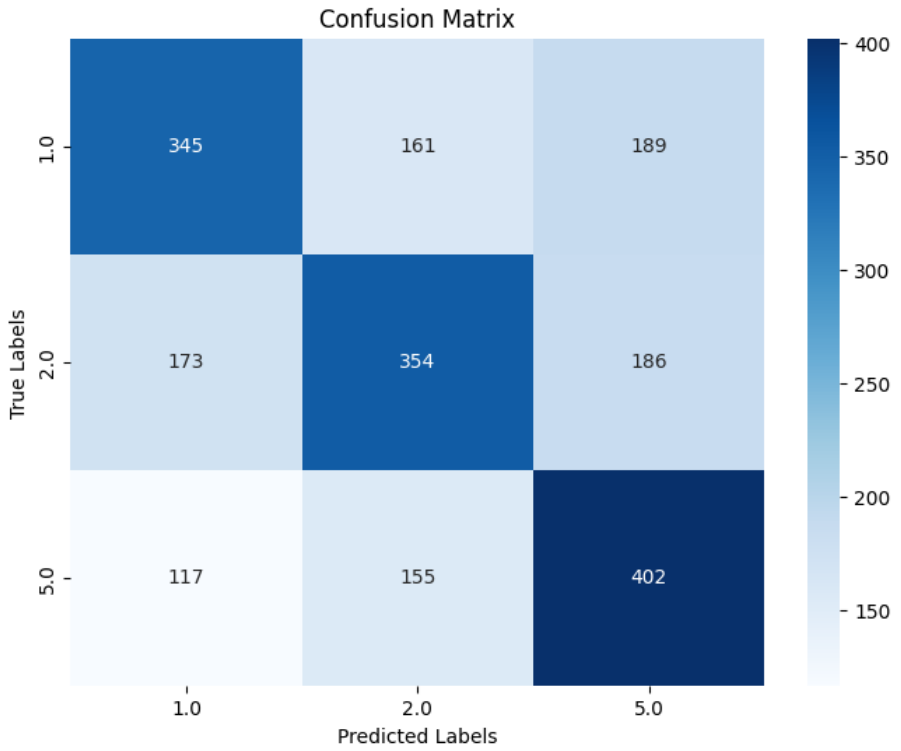

Figura 8. Confusion Matrix, SFS - Random Forest - SMOTE

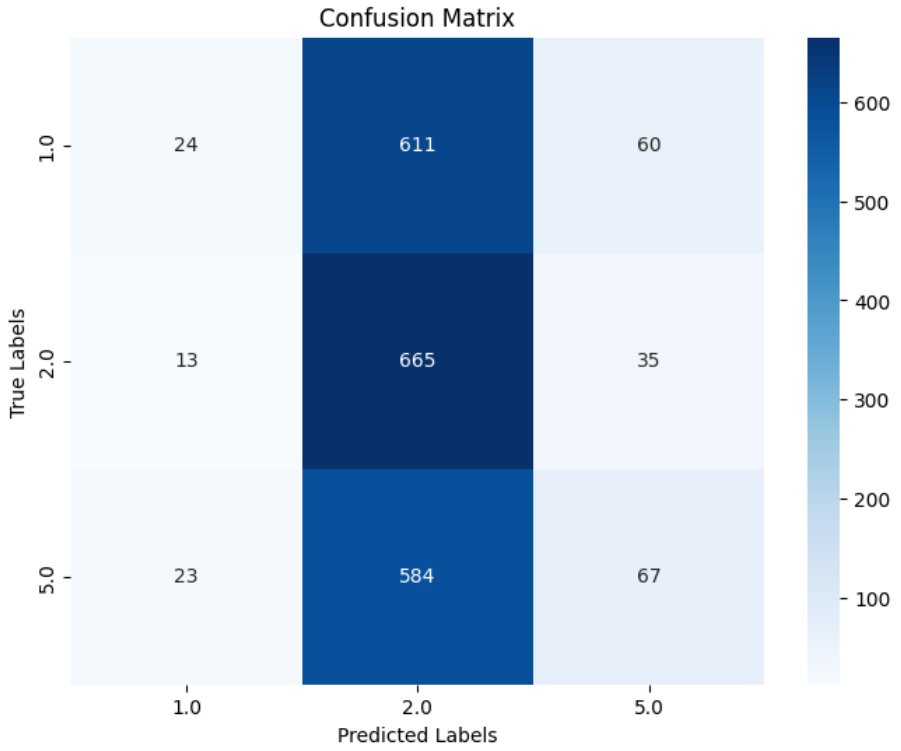

Figura 9. Confusion Matrix, SFS - Naive Bayes - SMOTE

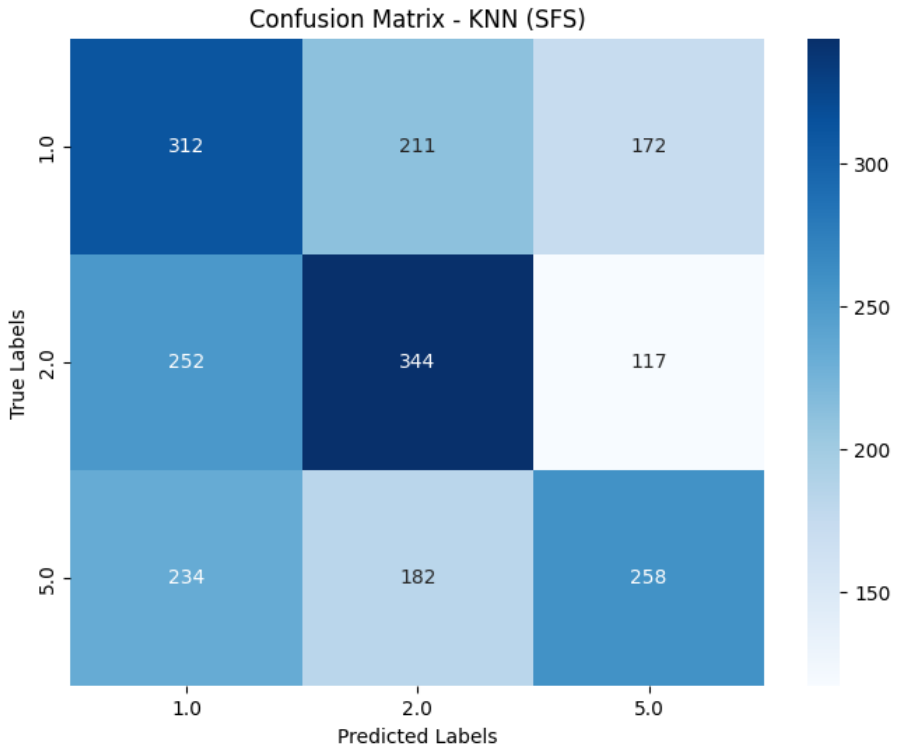

Figura 10. Confusion Matrix, SFS - KNN - SMOTE

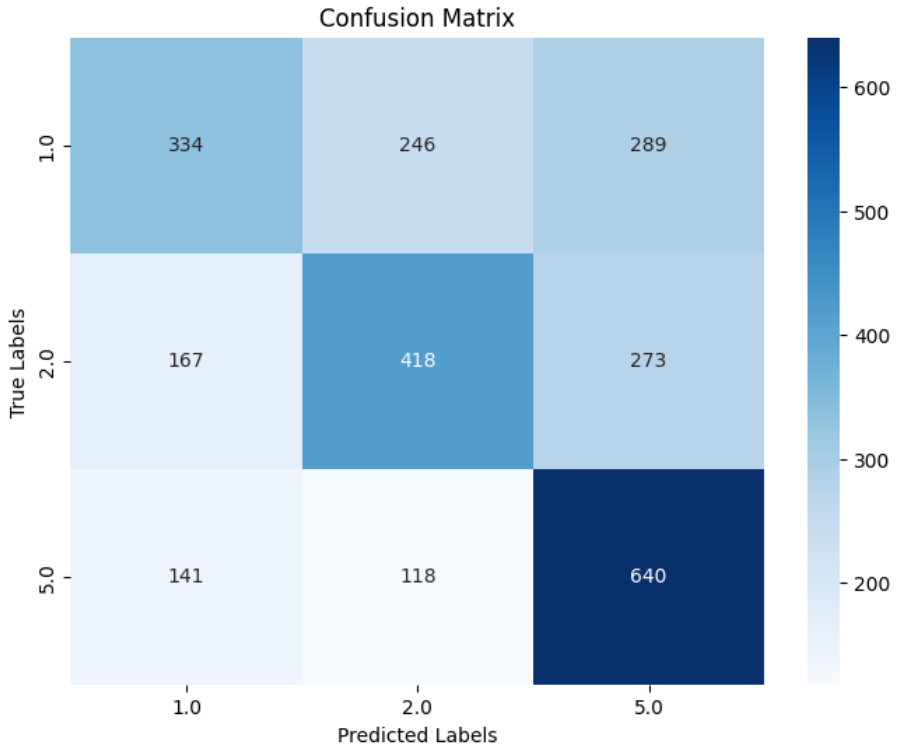

Figura 11. Confusion Matrix, SFS - Random Forest - Undersampling

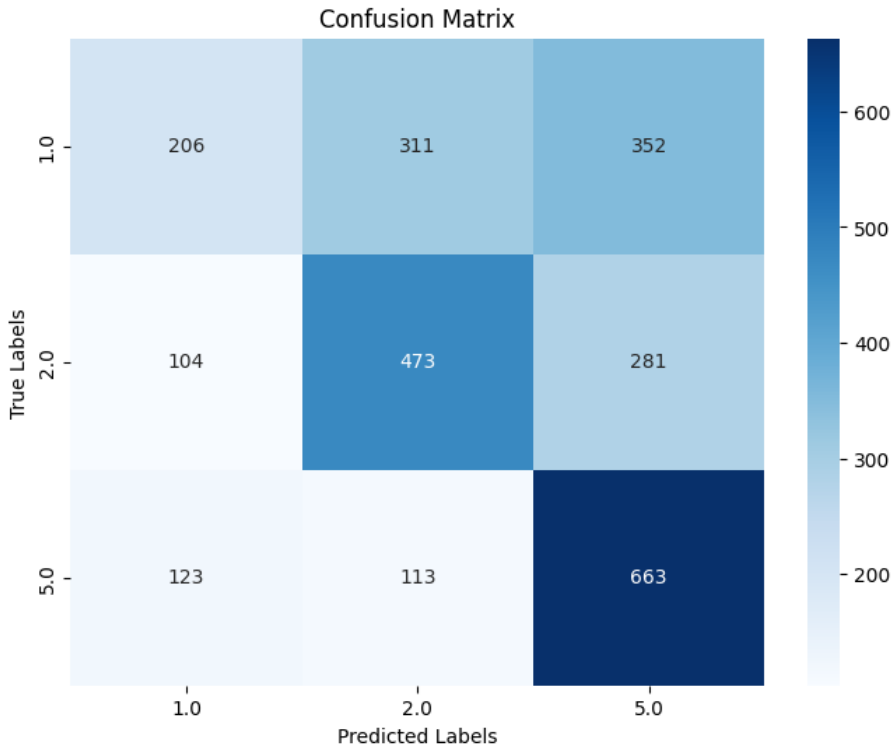

Figura 12. Confusion Matrix, SFS - Naive Bayes - Undersampling

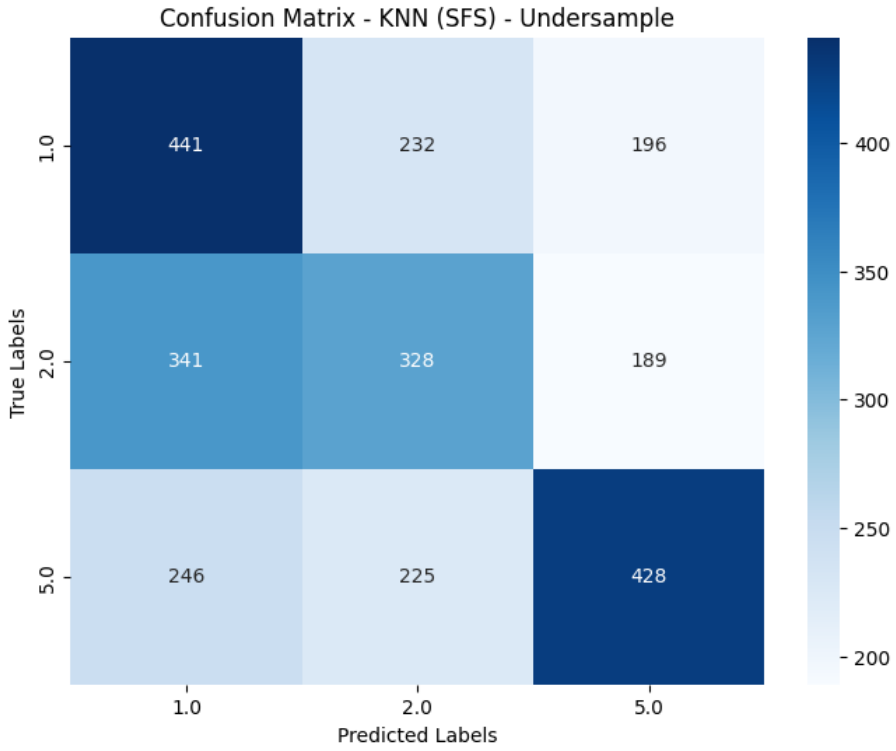

Figura 13. Confusion Matrix, SFS - KNN - Undersampling

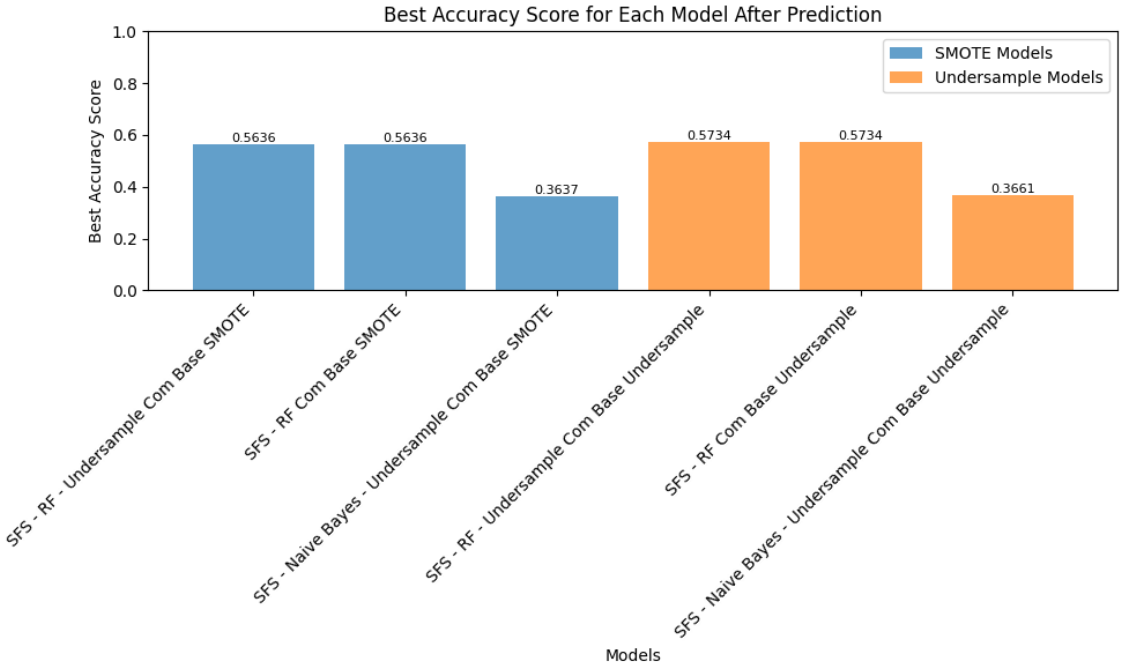

Figura 14. Treino Final com os 3 Modelos Finais entre as 2 Bases ”SMOTE”e ”Undersampling”

5. Conclusão

Diante da dificuldade de diagnóstico, tanto a influenza quanto a COVID-19 compartilham de muitos sintomas semelhantes, o que pode levar a diagnósticos paralelos ou retardo do tratamento correto e eficiente. Para ajudar a resolver esse problema, planejamos a integração das ML como medida de auxílio para a triagem e o diagnóstico. Entretanto enfrentamos limitações pelo alto desbalanceamento de dados.

Neste estudo, conduzimos a criação de seis modelos de Machine Learning, utilizando três algoritmos distintos (Random Forest, Naive Bayes e K-NN), aplicados a duas bases de dados balanceadas utilizando duas técnicas (SMOTE e Undersampling). A partir desses modelos, selecionamos os três melhores com base em métricas de acurácia, os quais foram posteriormente refinados por meio do processo de otimização de hiperparâmetros Random Search. Em seguida, realizamos testes desses três modelos nas bases de dados obtidas, avaliando suas métricas finais. O modelo que apresentou o melhor desempenho dentre os citados foi Random Forest, assim ajudando na identificação e classificação dos casos de influenza, COVID-19 e outros Nossos resultados mostram uma boa classificação. Entretanto, devido às limitações, obtivemos os diagnósticos em estatística.

5.1. Trabalhos futuros

O levantamento de dados foi uma etapa difícil durante a realização deste trabalho, em função da indisponibilidade de algumas informações e do tempo para a conclusão desta pesquisa. Recomenda-se que, para os trabalhos futuros, haja diversificação das fontes de dados: integrar dados das unidades de saúde particulares, clínicas e dados globais, para o desenvolvimento de uma avaliação mais precisa e abrangente, trazendo assim a possibilidade de uma avaliação por risco. Para a finalização das implementações e integrações, recomenda-se o desenvolvimento de APIs e interfaces de usuário, para que o usuário final possa realizar a triagem e obter o diagnóstico final. Também possibilitamos a ideia de que, no futuro, o projeto possa utilizar a análise de séries temporais para monitorar tendências e padrões de incidência de doenças, a fim de alertar possíveis surtos epidemiológicos.

Referências

BATISTA, G. E. d. A. P. et al. *Pré-processamento de dados em aprendizado de máquina supervisionado*. Tese (Doutorado) — Universidade de São Paulo, 2003. 4

CHAPMAN, P. et al. The crisp-dm user guide. In: SN. *4th CRISP-DM SIG Workshop in Brussels in March*. [S.l.], 1999. v. 1999. 7

Dados.gov.br. *SRAG - 2009-2012*. s.d. Acessado em: [06/06/2024]. Disponível em: ⟨https://dados.gov.br/dados/conjuntos-dados/srag-2009-2012⟩. 2

D'AMATO, G. et al. Covid-19, air pollution, and climate change. *Journal of Allergy and Clinical Immunology*, v. 146, n. 4, p. 725–727, 2020. Acessado em: [09/06/2024]. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0883944120306237⟩. 6

DataGeeks. *K-Nearest Neighbors (KNN)*. 2021. Acessado em: [09/06/2024]. Disponível em: ⟨https://www.datageeks.com.br/k-nearest-neighbors/⟩. 5

EBAC Online. *Random Forest para SEO: Entenda como essa técnica pode revolucionar suas estratégias*. 2021. Acessado em: [09/06/2024]. Disponível em: ⟨https://ebaconline.com.br/blog/random-forest-seo⟩. 5

Insper. *Mercado de Inteligência Artificial cresce cada vez mais acelerado*. s.d. Acessado em: [09/06/2024]. Disponível em: ⟨https://www.insper.edu.br/noticias/mercado-de-inteligencia-artificial-cresce-cada-vez-mais-acelerado/⟩. 3

LI, Y. et al. Effects of molecular weight on the antibacterial activity and mechanism of action of chitosan against gram-negative bacteria. *Scientific Reports*, v. 11, n. 1, p. 1–12, 2021. Acessado em: [09/06/2024]. Disponível em: ⟨https://www.nature.com/articles/s41598-021-93832-2⟩. 6

MAGHDID, H. S. et al. A novel ai-enabled framework to diagnose coronavirus covid-19 using smartphone embedded sensors: design study. *npj Digital Medicine*, v. 3, n. 1, p. 1–11, 2020. Acessado em: [09/06/2024]. Disponível em: ⟨https://www.nature.com/articles/s41746-020-00372-6⟩. 6

Ministério da Saúde. *Dicionário de Dados SRAG Hospitalizado*. 2022. PDF. Acessado em: [09/06/2024]. Disponível em: ⟨https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SRAG/pdfs/Dicionario_de_Dados_SRAG_Hospitalizado_19.09.2022.pdf⟩. 8

MOHAMMED, R.; RAWASHDEH, J.; ABDULLAH, M. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In: IEEE. *2020 11th international conference on information and communication systems (ICICS)*. [S.l.], 2020. p. 243–248. 5

SANTIAGO, D. *Estratégias eficazes para lidar com conjuntos de dados desbalanceados*. 2021. Acessado em: [05/06/2024]. Disponível em: ⟨https://medium.com/@daniele.santiago/estrat%C3%A9gias-eficazes-para-lidar-com-conjuntos-de-dados-desbalanceados-5b873894483b⟩. 4

Secretaria de Estado da Saúde de Santa Catarina. *Síndrome Respiratória Aguda Grave (SRAG)*. s.d. Acessado em: [07/06/2024]. Disponível em: ⟨https://dive.sc.gov.br/index.php/sindrome-respiratoria-aguda-grave-srag⟩. 2

SRAG 2021 e 2022. 2022. Dados Abertos. Acessado em: [10/05/2024]. Disponível em: ⟨https://dados.gov.br/dados/conjuntos-dados/srag-2021-e-2022⟩. 9

SRAG, P. *Projeto SRAG*. 2024. ⟨https://github.com/LGFFProj/Projeto_SRAG⟩. Repositorio do Projeto. 14

VALANDRO, E. *Feature Selection: A chave para modelos de machine learning mais eficientes*. 2021. Acessado em: [09/06/2024]. Disponível em: ⟨https://pt.linkedin.com/pulse/feature-selection-chave-para-modelos-de-machine-mais-e-valandro-prv2f⟩. 5

Varella, Drauzio. *Gripe H1N1 (Gripe Suína)*. s.d. Acessado em: [09/06/2024]. Disponível em: ⟨https://drauziovarella.uol.com.br/doencas-e-sintomas/gripe-h1n1-gripe-suina/⟩. 2

VARGAS, C. L. et al. Diagnóstico laboratorial da covid-19: uma revisão de técnicas moleculares. *Jornal Brasileiro de Patologia e Medicina Laboratorial*, v. 58, n. 1, p. 1–10, 2022. Disponível em: ⟨https://www.scielo.br/j/jbpml/a/zFfHzH4zZ48wWtPVWxzzjbc/?lang=pt⟩. 2

VARGAS, C. L. et al. Diagnóstico laboratorial da covid-19: uma revisão de técnicas moleculares. *Jornal Brasileiro de Patologia e Medicina Laboratorial*, v. 58, n. 1, p. 1–10, 2022. Acessado em: [09/06/2024]. Disponível em: ⟨https://www.scielo.br/j/jbpml/a/zFfHzH4zZ48wWtPVWxzzjbc/?lang=pt⟩. 6

WANG, J. et al. Classification of imbalanced data by using the smote algorithm and locally linear embedding. In: IEEE. *2006 8th international Conference on Signal Processing*. [S.l.], 2006. v. 3. 4

Wikipedia. *Teorema de Bayes*. 2021. Acessado em: [09/06/2024]. Disponível em: ⟨https://pt.wikipedia.org/wiki/Teorema_de_Bayes⟩. 5

YIN, J. et al. Machine learning for invasive fungal disease diagnosis using ct imaging: A pilot study. *Heliyon*, v. 10, n. 4, p. e041744, 2024. Acessado em: [09/06/2024]. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S2405844024041744⟩. 6