Análise Comparativa de Modelos de Machine Learning para Detecção Precoce de Câncer de Pulmão

Thales Henrique Bastos Neves
Departamento de Computação
CEFET-MG
Belo Horizonte, Minas Gerais
thaleshenrique44@gmail.com

Fábio Rocha da Silva
Departamento de Computação
CEFET-MG
Belo Horizonte, Minas Gerais
fabiorochadasilva@cefetmg.br

Luiz Fernando Pinheiro Ramos
Departamento de Computação
CEFET-MG
Belo Horizonte, Minas Gerais
luizramos@cefetmg.br

***Resumo*—O câncer de pulmão representa um dos maiores desafios para a saúde pública, sendo responsável por elevadas taxas de mortalidade devido ao diagnóstico tardio e à limitação no acesso a exames especializados. Diante desse cenário, este trabalho investigou o potencial do aprendizado de máquina como ferramenta de apoio à detecção precoce da doença, a partir de dados clínicos estruturados. Embora existam avanços na literatura com modelos preditivos para câncer, muitos ainda carecem de validação em contextos realistas e apresentam limitações quanto à interpretabilidade e abrangência. Assim, propôs-se um estudo comparativo entre sete algoritmos de classificação supervisionada, avaliando sua eficácia em prever casos de câncer de pulmão com base em sintomas clínicos. A metodologia envolveu o pré-processamento dos dados, balanceamento com a técnica SMOTE[1] e avaliação dos modelos por métricas como acurácia, precisão, *revocação (recall)*, *Área sob a Curva (AUC)* e f1-score. O modelo Random Forest apresentou o melhor desempenho geral, alcançando a acurácia de 92,77%, *AUC* de 0,991 e *recall* de 100 %, o que o torna especialmente adequado para o contexto clínico por maximizar a detecção de casos positivos. Modelos como *K-Nearest Neighbors* (*KNN*) e *XGBoost* também obtiveram boas métricas de desempenho, mas com sensibilidade ligeiramente inferior. O trabalho reforça, assim, o potencial da inteligência artificial como ferramenta auxiliar na tomada de decisões médicas, promovendo diagnósticos mais rápidos, precisos e embasados em dados.**

Palavras-chave: câncer de pulmão; aprendizado de máquina; diagnóstico precoce; saúde pública; modelos preditivos; *SVM*; *random forest*; *KNN*; *XGBoost*.

I. Introdução

Os casos de câncer de pulmão no Brasil devem crescer 65% e a mortalidade 74% até 2040, se mantido o atual padrão de consumo de tabaco no país (estimativa alarmante da Fundação do Câncer com base em dados da Agência Internacional de Pesquisa sobre o Câncer. Essa projeção revela um desafio crítico para a saúde pública: a doença persiste como uma das neoplasias mais letais, com o Instituto Nacional de Câncer (INCA) projetando 14 mil novos casos em mulheres e 18 mil em homens apenas em 2024. O problema se agrava quando observamos as disparidades regionais: enquanto a média nacional é de 12 casos por 100 mil habitantes (homens), o Sul do país registra o dobro dessa incidência. Esses números não apenas refletem a urgência do tema, mas apontam para uma lacuna crucial: a necessidade de estratégias inovadoras para superar as limitações do diagnóstico tardio e das desigualdades no acesso à saúde. Neste contexto, a aplicação de técnicas de análise de dados e aprendizado de máquina surge como alternativa promissora para identificar padrões clínicos preditivos [1][2] [2].

O câncer de pulmão se destaca como uma das neoplasias mais incidentes e letais no Brasil, refletindo um cenário preocupante para a saúde pública nacional. Apesar dos avanços nas políticas antitabagismo e na redução da prevalência do fumo, a doença continua a apresentar altas taxas de mortalidade, especialmente devido ao diagnóstico tardio e às disparidades no acesso aos serviços de saúde. Fatores como o desconhecimento dos sintomas por parte dos pacientes, limitações na infraestrutura diagnóstica e desigualdades entre os sistemas público e privado de saúde agravam a situação. Estudo indicam que aproximadamente 70% dos casos são diagnosticados em estágios avançados, comprometendo significativamente as chances de cura. Além disso, a taxa de sobrevida em cinco anos no país permanece em torno de 18%, similar às médias globais [3].

Diante da elevada mortalidade associada ao câncer de pulmão e das dificuldades persistentes no diagnóstico precoce, torna-se essencial o uso de abordagens técnicas avançadas para compreender melhor os padrões epidemiológicos da doença. A análise de grandes volumes de dados clínicos, ambientais e demográficos por meio de métodos estatísticos e algoritmos de aprendizado de máquina permite identificar correlações relevantes, prever riscos e orientar políticas públicas mais eficazes. Técnicas como redes neurais, árvores de decisão e modelos de séries temporais têm se mostrado promissoras na detecção precoce e na análise de fatores associados ao agravamento da condição [4]. Essas ferramentas podem oferecer suporte

[1]SMOTE (Synthetic Minority Over-sampling Technique) é uma técnica de balanceamento que gera exemplos sintéticos da classe minoritária, ajudando o modelo a aprender melhor padrões em conjuntos de dados desbalanceados.

[2]Para acessar a primeira referência, certifique-se de que o link contém o hífen entre `radioagencia` e `nacional`, como em: `radioagencia-nacional`. Caso o hífen não apareça, copie e cole o link manualmente no navegador.

à tomada de decisão médica, otimizar recursos e, sobretudo, contribuir para estratégias mais assertivas de prevenção e controle do câncer de pulmão.

Inicialmente, este trabalho foi concebido com o objetivo de realizar uma análise aprofundada da relação entre a poluição do ar e a incidência de câncer de pulmão no estado de Minas Gerais, utilizando dados públicos nacionais. A proposta previa o uso do *DataSUS* para acessar informações sobre internações e diagnósticos da doença em nível estadual, cruzando esses dados com indicadores ambientais, oriundos de plataformas como o IBGE e o Instituto Nacional de Meteorologia (INMET). No entanto, ao longo do desenvolvimento do projeto, foram identificadas diversas dificuldades técnicas, como a fragmentação dos dados, inconsistências nos formatos, ausência de padronização temporal e espacial, além da limitação na integração entre os conjuntos de dados de saúde e ambientais. Essas limitações comprometeram a viabilidade da abordagem inicialmente proposta, especialmente dentro dos prazos acadêmicos disponíveis.

Com a impossibilidade de seguir com a proposta inicial envolvendo dados ambientais e de saúde de Minas Gerais, o foco do projeto foi redirecionado exclusivamente para a análise da incidência do câncer de pulmão com base nas informações clínicas e sintomatológicas presentes em uma base de dados internacional disponível na plataforma *Kaggle*. Essa base, amplamente utilizada em estudos acadêmicos e experimentos com aprendizado de máquina, contém dados estruturados sobre sintomas apresentados pelos pacientes, presença ou ausência da doença e outras características relevantes. A nova abordagem, mais acessível e tecnicamente viável, permitiu manter a proposta de aplicar técnicas de inteligência artificial para desenvolver modelos preditivos voltados à identificação de padrões associados ao diagnóstico da doença, contribuindo com *insights* importantes para a área da saúde.

A metodologia adotada neste trabalho envolve diversas etapas fundamentais para garantir a qualidade e a robustez das análises. Inicialmente, os dados foram submetidos ao pré-processamento, que incluiu a limpeza de valores ausentes ou inconsistentes. Em seguida, foi realizada a separação da base em conjuntos de treino e teste, além do balanceamento das classes, permitindo a validação adequada dos modelos. Diversas técnicas de aprendizado de máquina foram aplicadas, como regressão logística, *random forest* e redes neurais artificiais, visando identificar quais sintomas e características clínicas apresentam maior influência sobre a presença do câncer de pulmão. Além disso, foram utilizadas métricas de desempenho, como acurácia, precisão, sensibilidade e área sob a curva *ROC*, para avaliar a eficácia dos modelos propostos. Essa abordagem permitiu não apenas a criação de classificadores eficientes, mas também uma compreensão mais clara dos padrões associados à doença.

II. Trabalhos Relacionados

Diversos estudos têm explorado o uso de técnicas de aprendizado de máquina na predição do câncer de pulmão, visando aprimorar o diagnóstico precoce e a tomada de decisões clínicas. Alsinglawi et al. (2022) propuseram um modelo preditivo para estimar o tempo de permanência hospitalar (*Length of Stay* – LOS) de pacientes com câncer de pulmão, utilizando algoritmos de aprendizado supervisionado, como *Random Forest*, XGBoost e regressão logística, validados por meio de uma abordagem de validação cruzada K-fold com 10 divisões. O estudo foi baseado no conjunto de dados *MIMIC-III*, que reúne informações de internações em unidades de terapia intensiva (UTI). Devido ao desbalanceamento das classes, os autores aplicaram técnicas de *oversampling*, especificamente o método Synthetic Minority Oversampling Technique (SMOTE), para equilibrar os dados durante o processo de validação. A pesquisa analisou um total de 53.423 pacientes adultos, e os resultados indicaram que o modelo *Random Forest*, aliado à técnica de balanceamento SMOTE, apresentou o melhor desempenho em comparação com os demais classificadores, alcançando uma *AUC* de 98% (intervalo de 95,3% a 100%) e uma taxa de revocação (*recall*) também de 98% (entre 95,3% e 100%) [5] [6].

Wu et al (2019). propuseram um modelo de *Random Forest* para a identificação de câncer de pulmão com base em índices hematológicos rotineiros. O desempenho do modelo foi avaliado utilizando validação cruzada com 10 subconjuntos (*10-fold cross-validation*), sobre um conjunto de dados contendo 277 pacientes coletados na Universidade de Lanzhou. Os resultados reportados foram expressivos, com acurácia de 95,7%, *recall* de 96,3% e área sob a curva *ROC* (*AUC*) de 99,01% [7].

Já o Vikas (2021) utilizou dois algoritmos de aprendizado de máquina, *Support Vector Machine* (SVM) e *Random Forest*, para prever casos de câncer de pulmão. Os autores realizaram comparações com e sem a aplicação do método de seleção de atributos *Chi-square*, observando que o modelo baseado em *SVM* apresentou melhor desempenho em termos de acurácia e tempo de execução. O modelo atingiu uma acurácia de 98%, precisão de 100%, *recall* de 100% e *F1-score* de 100%, com tempo de execução de apenas 0,010 segundos. O conjunto de dados utilizado foi obtido no site "Data World", contendo 1000 amostras e 25 atributos [8]. Para o mesmo conjunto de dados P.R., Radhika, Nair e Gangadharan (2019) realizaram um estudo para prever o câncer de pulmão utilizando quatro algoritmos de aprendizado de máquina: *Naive Bayes*, *Support Vector Machine* (SVM), *Árvore de Decisão* e *Regressão Logística*. O objetivo principal da pesquisa foi promover o diagnóstico precoce da doença e analisar comparativamente o desempenho desses modelos. Dentre os algoritmos testados, o *Support Vector Machine* obteve o melhor resultado, com uma acurácia de 99,2% [9].

Por fim, Kononenko (2021) argumentou que a aprendizagem de máquina é um dos principais ramos da inteligência artificial e tem se mostrado uma ferramenta indispensável na análise inteligente de dados médicos [10]. Desde seus primeiros usos, os algoritmos de aprendizado de máquina (*machine learning*) foram projetados para analisar conjuntos de dados clínicos, e hoje sua aplicação se estende a diversas áreas da medicina, incluindo o diagnóstico de doenças complexas. O autor

destaca que, especialmente com o avanço das tecnologias de informação e a digitalização dos sistemas hospitalares, tornou-se possível coletar grandes volumes de dados clínicos com facilidade. Esses dados, ao serem processados por algoritmos de aprendizado, podem ser utilizados para construir classificadores capazes de auxiliar médicos no diagnóstico de novos pacientes, aumentando a precisão, a velocidade e a confiabilidade das decisões médicas. Além disso, esses modelos podem ser utilizados no treinamento de profissionais não especialistas em tarefas de diagnóstico específicas. O estudo também oferece uma visão histórica, atual e futura sobre o papel da inteligência artificial na medicina, enfatizando métodos como árvores de decisão, redes neurais e classificadores Bayesianos, e apontando para novas abordagens que consideram a confiabilidade das decisões dos modelos e a validação de fenômenos ainda não explicados pela medicina tradicional.

III. Metodologia

A metodologia adotada neste trabalho foi estruturada em etapas sequenciais, com foco em garantir a consistência dos dados, o bom desempenho dos modelos de aprendizado de máquina e a interpretabilidade dos resultados. Primeiro, foram coletados os dados de incidência de câncer de pulmão; em seguida, foi realizada a descrição dos atributos presentes na base de dados, com o objetivo de compreender melhor as variáveis disponíveis e sua relação com a presença da doença. Na próxima etapa, foi realizado o pré-processamento dos dados para identificação de padrões, balanceamento e limpeza. Posteriormente, desenvolveu-se os modelos de predição e, por fim, a avaliação de desempenho e a análise dos resultados para identificar os modelos que apresentaram os melhores resultados. Essas etapas da metodologia podem ser observadas no fluxograma da Figura 1.

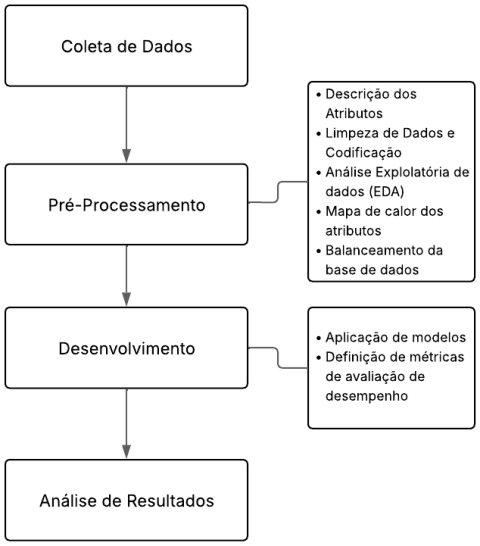

Figura 1: Fluxograma da metodologia

A. Coleta de Dados

A base de dados utilizada neste trabalho foi obtida a partir de um conjunto de dados disponível na plataforma Kaggle, intitulado *"Lung Cancer"* [11]. Esta base foi escolhida por sua acessibilidade e pela presença de variáveis clínicas e comportamentais diretamente relacionadas a fatores de risco do câncer de pulmão, o que viabiliza a análise preditiva com técnicas de aprendizado de máquina sem a necessidade de extensos processos de limpeza e integração de dados. O conjunto de dados é composto por informações de 309 indivíduos e apresenta como variável-alvo a presença ou ausência de câncer de pulmão, codificada como variável binária. Destaca-se que a maioria dos participantes da base de dados é composta por pessoas idosas, o que está alinhado com o perfil epidemiológico do câncer de pulmão. Porém, também há registros de pessoas mais jovens, o que permite uma análise mais abrangente dos fatores de risco. Vale acrescentar que o motivo da escolha dessa base de dados foi devido a facilidade e praticidade dela, diferente das bases de dados dos trabalhos relacionados como a *Data World* e *MIMIC-III*, que não foram encontradas.

Descrição de atributos: A descrição de atributos é essencial para entender melhor a sua base de dados, como ela esta estruturada e quais são os atributos utilizados nela. Foram utilizados os atributos mostrados na Tabela I:

Tabela I: Descrição dos atributos

Atributo	**Descrição**
Age	Idade do indivíduo participante (em anos).
Smoking	Indica se a pessoa é fumante.
Yellow_Fingers	Indica se o indivíduo tem dedos amarelados. Um sinal físico comum em fumantes crônicos.
Anxiety	Informa se o indivíduo apresenta sintomas de ansiedade.
Peer_Pressure	Informa se os indivíduos relataram sofrer influência da pressão de colegas em hábitos como fumar entre outros.
Chronic_Disease	Indica se o indivíduo possui alguma doença crônica, como diabetes ou hipertensão.
Fatigue	Indica se o indivíduo apresenta sinais de fadiga, cansaço ou exaustão frequente.
Allergy	Indica se o indivíduo apresenta algum tipo de alergia conhecida (respiratória, alimentar, etc.).
Wheezing	Indica se o indivíduo apresenta sintomas de chiado no peito, geralmente associado a problemas respiratórios.
Alcohol_Consuming	Indica se o indivíduo consome álcool regularmente.
Coughing	Indica a presença de tosse persistente, um sintoma comum em doenças respiratórias.
Shortness_of_Breath	Informa se o indivíduo sente dificuldade para respirar, principalmente durante esforço.
Swallowing_Difficulty	Indica se há presença de disfagia, ou seja, dificuldade para engolir alimentos ou líquidos.
Chest_Pain	Indica se o indivíduo sente dores na região torácica, podendo estar associada a problemas respiratórios ou cardíacos.

B. Pré-processamento de Dados

As seguintes etapas foram realizadas para garantir a qualidade dos dados antes da modelagem:

- **Limpeza de Dados**: verificação e remoção de dados ausentes, duplicados ou inconsistentes.
- **Balanceamento**: aplicação de técnicas de *oversampling* e *undersampling* para equilibrar a proporção entre classes (diagnóstico positivo e negativo), com o uso do SMOTE para balancear as classes antes do treinamento de cada modelo, pois o conjunto de dados utilizado é desbalanceado [6].

Análise Exploratória de Dados: A análise exploratória de dados (EDA) foi conduzida com o objetivo de compreender melhor o comportamento das variáveis, identificar padrões relevantes e potenciais relações entre os fatores de risco e o diagnóstico de câncer de pulmão. A escolha por analisar especificamente a razão de casos positivos deve-se ao objetivo central de identificar padrões associados à ocorrência da doença. Enquanto os casos negativos são importantes para estudos de especificidade ou comparação de grupos, a análise priorizou entender as características predominantes nos indivíduos diagnosticados, pois isso: (i) direciona a investigação para fatores de risco potencialmente críticos; (ii) evita diluir os sinais estatísticos com dados não relevantes para o fenômeno estudado. Essa abordagem está alinhada ao foco preventivo do estudo, que visa apoiar a identificação precoce e o encaminhamento de pacientes com maior risco para exames complementares e acompanhamento médico adequado, por isso foi feito a analise com foco nos casos positivos. As principais etapas e visualizações realizadas foram:

I. **Distribuição de Gêneros da Base de Dados**: a proporção de indivíduos por gênero foi analisada por meio de gráficos de pizza, revelando uma distribuição equilibrada entre homens (51,45%) e mulheres (48,55%). Essa balanceamento é importante, pois evita viés em relação a um dos gêneros e permite que os resultados obtidos pelos modelos possam ser generalizados de forma mais confiável para ambos os grupos.

II. **Distribuição de Idades**: foi gerado um gráfico para examinar a distribuição etária dos indivíduos diagnosticados com câncer de pulmão. A análise mostrou uma concentração maior de casos em idades acima dos 50 anos, confirmando a prevalência da doença em pessoas idosas, conforme a Figura 2.

III. **Razão de Casos Positivos por Gênero em Fumantes e Consumidores de Álcool**: foram gerados gráficos para comparar a taxa de casos positivos entre homens e mulheres com histórico de tabagismo e consumo de álcool. A análise destacou que, entre os consumidores de álcool, os homens apresentaram um maior valor absoluto de diagnósticos positivos, com 101 casos frente a 44 casos entre as mulheres. Entre os que não consumiam álcool, as mulheres superaram os homens, com 69 casos contra 24. Entre os fumantes, os padrões foram semelhantes, indicando o impacto combinado desses fatores comportamentais. Homens fumantes registraram 71 casos, enquanto os não fumantes tiveram 54. Já entre as mulheres, foram 60 casos entre fumantes e 53 entre não fumantes, como está mostrado a Figura 3. Esses números reforçam que tanto o tabagismo quanto o consumo de álcool estão associados a uma maior incidência de casos, especialmente entre os homens, só reforçando que nessa análise estamos observando apenas quem teve diagnóstico positivo de câncer de pulmão.

IV. **Distribuição de Sintomas para Casos Positivos em Termos de Gênero**: a análise dos sintomas (como tosse, chiado no peito, falta de ar, entre outros) foi segmentada por gênero nos casos positivos. Os gráficos de barras permitiram observar que certos sintomas estavam mais presentes em um dos sexos, oferecendo indícios de possíveis variações clínicas na manifestação da doença, esses gráficos podem ser vistos na Figura 4 e na Figura 5. Casos como dor no peito e tosse se destacaram entre os homens, com 94 e 85 ocorrências respectivamente, em contraste com 48 e 64 nas mulheres — uma diferença expressiva que sugere uma predominância desses sintomas no público masculino. O sintoma "chiado no peito"(*wheezing*) também foi significativamente mais relatado por homens (81 casos) do que por mulheres (61 casos). Já falta de ar (*short breath*) teve uma distribuição mais equilibrada: 80 casos em homens e 73 em mulheres. Por outro lado, sintomas como dedos amarelados (*yellow fingers*) apareceram com muito mais frequência nas mulheres (84 casos) do que nos homens (62 casos), indicando uma diferença marcante. O mesmo padrão foi observado em doenças crônicas (*chronic disease*), com 74 casos entre mulheres e 57 entre homens. Outros sintomas, como fadiga (84 homens vs. 81 mulheres), ansiedade (52 homens vs. 73 mulheres), alergia (85 homens vs. 61 mulheres), dificuldade para engolir (59 homens vs. 65 mulheres) e tosse (85 homens vs. 64 mulheres), mostraram variações mais moderadas, mas ainda relevantes clinicamente.

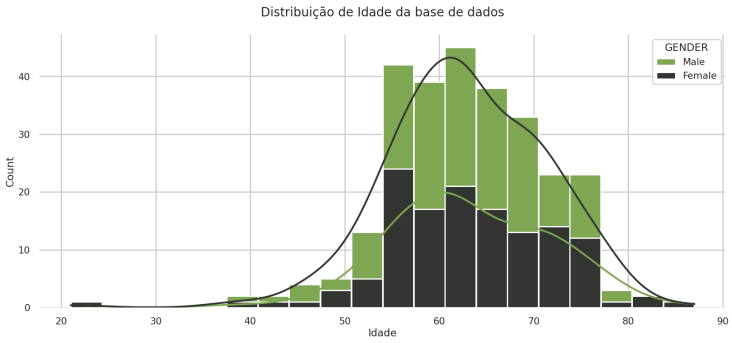

Figura 2: Gráfico da Distribuição de Idade

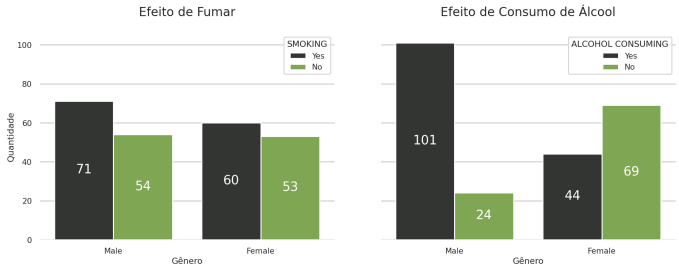

Figura 3: Gráfico de Consumo de Álcool e Fumantes para casos positivos

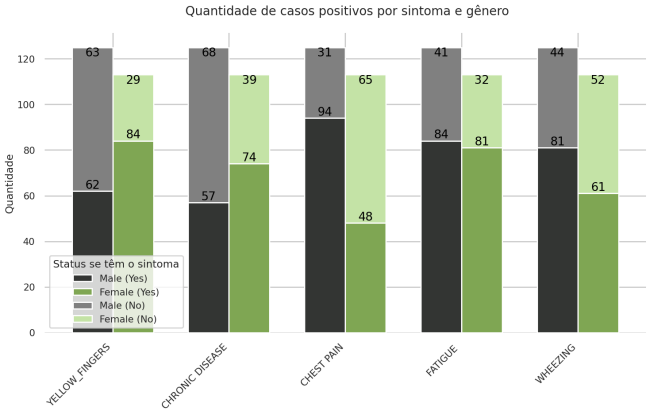

Figura 4: Gráfico 1 de Comparação dos Sintomas

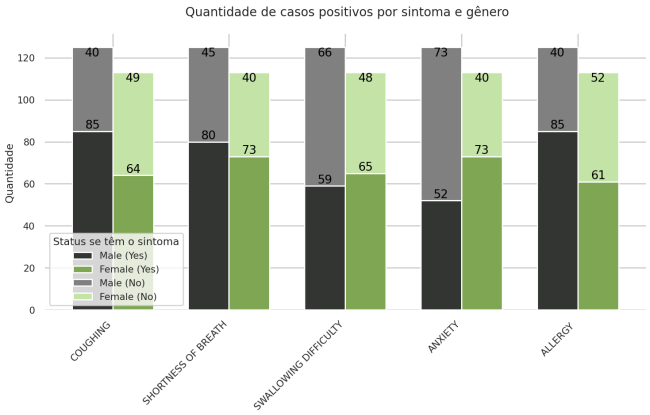

Figura 5: Gráfico 2 de Comparação dos Sintomas

C. Desenvolvimento

Para a implementação prática do projeto, utilizou-se a linguagem de programação *Python*, juntamente com diversas bibliotecas especializadas. As bibliotecas *Pandas*, *NumPy*, *Seaborn* e *Matplotlib* foram essenciais para o tratamento, manipulação e visualização dos dados.[3] Já o *Scikit-learn* teve um papel central na construção e avaliação dos modelos de aprendizado de máquina, enquanto o SMOTE foi utilizado para o balanceamento das classes e o XGBoost para a aplicação de algoritmos baseados em *boosting*.

Construção dos Modelos: Para a etapa de modelagem preditiva, foram selecionados algoritmos de aprendizado de máquina supervisionado amplamente utilizados em tarefas de classificação binária. O objetivo foi prever a ocorrência de câncer de pulmão com base em variáveis clínicas e comportamentais dos indivíduos da base de dados.

Procedimentos Comuns:

- Os dados foram divididos em dois subconjuntos: 70% para treinamento e 30% para teste.
- Balanceamento usando a técnica SMOTE para balancear a nossa base de dados utilizada na aplicação de cada modelo.
- As métricas de avaliação incluíram acurácia, precisão, *recall*, *F1-score*, *AUC* (Área sob a Curva ROC).

A seguir, são descritos os modelos aplicados e os procedimentos adotados:

I. **Regressão Logística**: a regressão logística é um modelo de aprendizado supervisionado utilizado para problemas de classificação binária. O modelo aprende os pesos dos atributos durante o treinamento, buscando minimizar a função de perda, o que o torna bastante eficiente e interpretável. A regressão foi empregada como um modelo base por sua simplicidade e interpretabilidade. Este modelo de regressão estima a probabilidade de um indivíduo ser diagnosticado com câncer de pulmão com base em uma combinação linear dos atributos de entrada, utilizando a função logística como ativação. [18].

II. **Naive Bayes Gaussiano**: o modelo de Naive Bayes Gaussiano assume que os dados seguem uma distribuição normal para cada classe. Neste trabalho, foi treinado com os dados balanceados e, em seguida, aplicado sobre o conjunto de teste para realizar previsões. Foi aplicado principalmente às variáveis contínuas, sendo útil por sua simplicidade e rapidez de treinamento, especialmente em conjuntos de dados pequenos como o presente. Essa abordagem permitiu avaliar a eficácia do classificador probabilístico em identificar corretamente os casos de câncer de pulmão. O modelo estima a probabilidade de cada classe, o que o torna adequado para variáveis contínuas como as utilizadas neste estudo. [19] [20].

III. **Naive Bayes de Bernoulli**: este modelo é especialmente adequado para atributos binários (0 ou 1), como a presença ou ausência de sintomas e hábitos de risco — características comuns no conjunto de dados analisado neste trabalho. Diferente do Naive Bayes Gaussiano, que assume uma distribuição normal das variáveis contínuas, o Bernoulli Naive Bayes avalia a probabilidade de ocorrência de eventos binários, sendo mais apropriado para representar respostas discretas.
[19] [20].

[3]As bibliotecas utilizadas no desenvolvimento deste trabalho incluem: *Scikit-learn* [12], *Matplotlib* [13], *Pandas* [14], *NumPy* [15], *SMOTE* [16] e *XGBoost* [17].

IV. **Support Vector Machine (SVM)**: o algoritmo SVM é um classificador supervisionado que busca encontrar o hiperplano ótimo que melhor separa as classes no espaço de atributos, maximizando a margem entre os dados de classes distintas. Essa margem é determinada pelos chamados *vetores de suporte* — os pontos de dados mais próximos ao limite de decisão [19] [21]. Neste trabalho, o modelo SVM foi aplicado com o *kernel* radial (*RBF*), adequado para capturar relações não lineares nos dados. Foram utilizados hiperparâmetros ajustados previamente, com $C = 100$ (penalização por erro de classificação) e $\gamma = 0{,}002$ (gamma = controle da influência de cada ponto de treino).

V. **Random Forest**: o modelo Random Forest consiste em um conjunto de árvores de decisão treinadas sobre subconjuntos aleatórios dos dados e atributos. A previsão final é obtida por votação majoritária entre as árvores, o que contribui para maior robustez, melhor generalização e redução da variância em relação a modelos individuais. Neste trabalho, foi utilizada uma floresta com 100 estimadores (árvores), fixando a semente aleatória para garantir reprodutibilidade dos resultados.

VI. **K-Nearest Neighbors (KNN)**: esse modelo classifica uma nova amostra com base na maioria das classes de seus k vizinhos mais próximos no espaço das características. O valor de k = 2 foi ajustado empiricamente, e a distância euclidiana foi utilizada como métrica de proximidade [18] [19]. O modelo foi treinado com os dados balanceados, uma vez que o desempenho do KNN é fortemente impactado pela escala das variáveis.

VII. **Gradient Boosting (XGBoost)**: é um algoritmo de aprendizado de máquina baseado em *boosting*, que constrói modelos sequenciais de árvores de decisão, em que cada nova árvore é treinada para corrigir os erros residuais das árvores anteriores. Essa abordagem permite uma modelagem mais robusta e precisa, sendo especialmente eficaz em conjuntos de dados tabulares e problemas complexos de classificação [22]. Apesar de não ter sido aplicada uma busca em grade (*grid search*) nesta etapa, o XGBoost já apresentou bom desempenho com os hiperparâmetros iniciais.

D. Avaliação de Desempenho

A avaliação de desempenho dos modelos preditivos será realizada com base em um conjunto de métricas amplamente utilizadas na literatura de aprendizado de máquina e análise de dados, as quais fornecem uma visão completa da eficácia dos modelos. As principais métricas de desempenho incluem:

- **Acurácia**: mede a proporção de previsões corretas em relação ao total de previsões feitas. Embora útil, não será a única métrica a ser considerada, uma vez que em problemas desbalanceados, como o nosso, ela pode ser enganosa.
- **Precisão e** *Recall*: a precisão mede a proporção de previsões positivas corretas entre todas as previsões positivas feitas, enquanto o *recall* avalia a proporção de previsões positivas corretas entre todos os casos positivos reais. Ambas são essenciais para avaliar a capacidade do modelo em identificar corretamente os casos de câncer de pulmão em situações desbalanceadas.
- **F1-Score**: o F1-Score combina precisão e *recall* em uma única métrica, proporcionando uma avaliação mais equilibrada do desempenho do modelo, especialmente quando há um trade-off entre essas duas métricas.
- ***AUC* (Área sob a Curva ROC)**: quantifica a capacidade do modelo em distinguir entre as classes positiva (com câncer) e negativa (sem câncer) em diferentes limiares de decisão. Um valor de *AUC* próximo de 1 indica excelente desempenho na separação das classes, sendo particularmente útil em contextos clínicos onde é importante avaliar a sensibilidade e especificidade do modelo simultaneamente.
- ***ROC* (Receiver Operating Characteristic)**: a avaliação dos modelos também incluiu a curva *ROC*. A *ROC* é uma ferramenta essencial para comparar classificadores binários em múltiplos limiares. As curvas e seus respectivos valores encontram-se na Figura 6.

IV. Análise dos Resultados

A análise dos resultados representa uma etapa fundamental do projeto, pois permite avaliar a eficácia dos algoritmos de aprendizado de máquina aplicados ao diagnóstico de câncer de pulmão. Nesta seção, são apresentados e discutidos os impactos do pré-processamento dos dados, a performance dos diferentes modelos treinados, bem como a interpretação das curvas *ROC* e valores de *AUC*.

Cada análise realizada — desde o balanceamento dos dados até a comparação das métricas — é essencial para compreender o comportamento dos modelos diante do problema proposto. Além de identificar qual algoritmo apresenta melhor desempenho, essa etapa permite avaliar a robustez, generalização e adequação clínica de cada abordagem, fornecendo subsídios para sua aplicação prática em cenários médicos reais.

A. Pré-processamento e Balanceamento dos dados

Inicialmente, os dados apresentavam uma distribuição desbalanceada, com 238 amostras de casos positivos de câncer (86,2% do total) e apenas 38 amostras negativas (sem câncer) (13,8% do total), o que poderia comprometer o desempenho dos modelos de classificação. Antes da modelagem, foi realizado o balanceamento da base de dados utilizando a técnica SMOTE, que aumentou o número de amostras da classe minoritária (sem câncer) para igualar o número de amostras da classe majoritária (com câncer). Com isso, obteve-se um conjunto de treino balanceado com 336 amostras, sendo 168 de cada classe. Essa etapa é crucial para evitar o viés do modelo em prever apenas a classe mais frequente. [6]

B. Desempenho dos Modelos

Foram utilizados sete modelos de classificação: KNN, XGBoost, Random Forest, Gaussian Naive Bayes, Bernoulli Naive Bayes, SVM e Regressão Logística. As métricas de

avaliação incluíram *acurácia*, *f1-score*, *precisão*, *revocação (recall)*, *Área sob a Curva ROC (AUC)* e análise visual por meio das curvas *ROC*, o resultado dessas métricas podem ser observadas na Figura 6 e na Tabela II.

Modelos de Melhor Desempenho : Os modelos *K-Nearest Neighbors* (KNN) e *Extreme Gradient Boosting* (XGBoost) lideraram em desempenho geral, ambos com acurácia de 95,18 % e *f1-score* de 97 %. O XGBoost atingiu *recall* perfeito (100 %), enquanto o KNN obteve 94 %, evidenciando sua alta capacidade de detectar pacientes com câncer de pulmão e, ao mesmo tempo, manter elevada taxa de acertos na classe negativa. Além disso, ambos registraram *AUC* de 0,985, indicando excelente separação entre as classes (Figura 6 e Tabela II). Essa combinação *recall* e *AUC* alto reduz significativamente o risco de falsos negativos — aspecto crítico no contexto clínico.

Modelos com Desempenho Competitivo: O *Random Forest* apresentou acurácia de 92,77 % e *f1-score* de 96 %, mas se destacou principalmente pelo *recall* de 100 % para a classe positiva e pelo maior *AUC* entre todos os modelos (0,991). Esses resultados demonstram sua robustez na detecção de casos de câncer, embora o valor de precisão (92 %) indique uma quantidade ligeiramente maior de falsos positivos em comparação a KNN e XGBoost.

Modelos Baseados em Naive Bayes: O *Gaussian Naive Bayes* alcançou acurácia de 92,77 %, *f1-score* de 96 % e *AUC* de 0,897, enquanto o *Bernoulli Naive Bayes* obteve acurácia de 90,36 % e *AUC* de 0,943. Ambos superaram o SVM em precisão (97–98 %), mas ainda carregam a limitação inerente de assumir independência entre as variáveis, o que pode restringir o aprendizado de correlações mais complexas.

Support Vector Machine (SVM): O SVM registrou acurácia de 90,36 %, *f1-score* de 94 %, precisão de 97 % e *recall* de 91 %, com um *AUC* expressivo de 0,967. Embora ligeiramente inferior aos melhores modelos em acurácia, seu alto *AUC* demonstra boa capacidade discriminativa, reforçando a utilidade do SVM em cenários de alta dimensionalidade.

Modelo de Regressão Logística: A Regressão Logística apresentou o menor desempenho entre os modelos analisados, com acurácia de 86,75 %, *f1-score* de 92 % e *AUC* de 0,866. Apesar de sua simplicidade e interpretabilidade, o menor *recall* (87 %) evidencia limitação na detecção de casos positivos, tornando-a menos indicada para aplicações médicas onde falsos negativos devem ser minimizados.

O *Random Forest* apresentou o melhor desempenho, com *AUC* de 0,991, refletindo excelente sensibilidade em praticamente toda a faixa de limiares e confirmando sua capacidade de detectar todos os casos de câncer (100 % de *recall*). Em seguida, KNN e XGBoost empataram com *AUC* de 0,985, evidenciando alta separação entre as classes, alinhada à elevada acurácia desses modelos.

O SVM (0,967) e a Regressão Logística (0,966) apresentaram desempenho muito próximo, demonstrando que, mesmo com fronteiras lineares ou kernel de margem máxima, é possível obter discriminação robusta. O Bernoulli Naive Bayes, com *AUC* de 0,943, manteve boa capacidade discriminativa, ao passo que o Gaussian Naive Bayes registrou 0,897, indicando ligeira perda de sensibilidade em limiares extremos, mostrando que esse modelo realmente não seria bom para o nosso problema pois a maioria de seus atributos são discretos, e o Naive Bayes Gaussiano tem como objetivo ser usado em atributos contínuos, então ele não se encaixa na nossa abordagem.

Em síntese, a análise das curvas *ROC* confirma a superioridade do Random Forest no critério de separação global, seguida de perto por KNN e XGBoost. Modelos lineares generalizados (Regressão Logística) e baseados em margens (SVM) mostraram-se competitivos, enquanto as abordagens Naive Bayes, embora simples, mantiveram desempenho aceitável. Esses resultados reforçam a importância de considerar a *AUC* em conjunto com métricas como *recall* e precisão, pois mesmo modelos com acurácia semelhante podem diferir significativamente em sua capacidade de classificação ao longo de diferentes limiares.

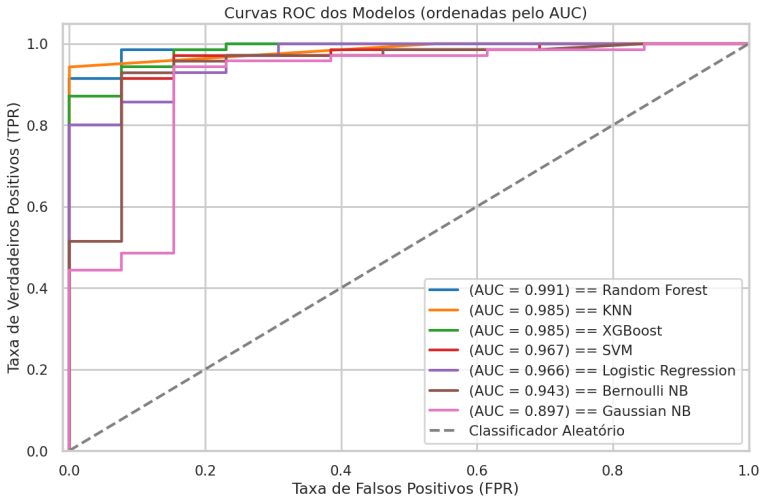

Figura 6: Curvas ROC

Tabela II: Desempenho dos Modelos de Classificação para Detecção de Câncer de Pulmão

Modelo	Acurácia	F1-Score	Precisão	Recall	AUC
KNN	95,18 %	97%	100%	94%	0,985
Bernoulli NB	90,36 %	94%	98%	90%	0,943
Gaussian NB	92,77 %	96%	97%	94%	0,897
SVM	90,36 %	94%	97%	91%	0,967
Logistic Regression	86,75 %	92%	97%	87%	0,966
XGBoost	95,18 %	97%	95%	100%	0,985
Random Forest	92,77 %	96%	92%	100%	0,991

C. Limitações da Análise Geral

Apesar dos resultados promissores, esta análise apresenta limitações importantes. O tamanho da base de dados é relativamente pequeno, o que pode impactar a capacidade de generalização dos modelos. Além disso, as variáveis utilizadas não contemplam fatores clínicos ou ambientais adicionais que poderiam influenciar o diagnóstico, como histórico familiar, exposição a poluentes ou tabagismo. Trabalhos futuros devem considerar conjuntos de dados maiores e de diferentes regiões, bem como validação cruzada mais robusta para assegurar maior confiabilidade dos modelos propostos.

D. Análise de Resultados por Gênero

Com o objetivo de avaliar a performance dos modelos em diferentes subgrupos populacionais, foi realizada uma análise separada por gênero, como mostrada na Tabela III e na Tabela IV. Essa abordagem permite identificar variações na sensibilidade e especificidade dos algoritmos, aspectos cruciais em contextos clínicos.

Entre os indivíduos do sexo masculino, o modelo *Random Forest* destacou-se com acurácia de 90,7%, *recall* de 92,1% e AUC de 0,958, indicando excelente capacidade de detecção dos casos positivos com baixo índice de falsos negativos. O *XGBoost*, embora com boa acurácia (90,7%), apresentou *recall* ligeiramente inferior (89,5%), AUC de 0,91 e acurácia de 90,7%.

Nos dados femininos, o melhor desempenho em termos de *recall* (93,3%) e AUC foi observado nos modelos *Regressão Logística* e *KNN*, ambos com acurácia de 92,68% e alta capacidade discriminativa (AUC superiores a 0,92). Esses modelos demonstraram ser mais adequados para identificar corretamente os casos positivos no grupo feminino. Em contraste, o *Random Forest*, que havia sido altamente eficaz no grupo masculino e no generalizado, apresentou menor acurácia (82,9%) e AUC de 0,91 nesse subconjunto. O *Gaussian Naive Bayes*, por sua vez, teve o pior desempenho entre as mulheres, com AUC de apenas 0,83 e *recall* de 83,3%, indicando limitação em identificar adequadamente as pacientes com câncer, o que faz sentido já que o modelo é especializado em variáveis contínuas, e a nossa base de dados utiliza variáveis discretas

Essas diferenças reforçam a importância do uso de múltiplas métricas para avaliação, em especial o *recall* e o AUC, que são cruciais para aplicações clínicas ao reduzir o risco de não detecção da doença. Também apontam para a necessidade de considerar a heterogeneidade populacional no desenvolvimento e validação dos modelos.

Cabe destacar como limitação dessa análise que a divisão da base de dados por gênero reduziu significativamente o número de amostras em cada grupo, o que limita a robustez estatística das comparações. Futuras pesquisas com conjuntos de dados maiores e mais balanceados entre os subgrupos poderão validar de forma mais consistente essas observações.

Tabela III: Métricas de desempenho dos modelos - Dados Masculinos

Modelo	**Acurácia**	**F1-Score**	**Precisão**	**Recall**	**AUC**
KNN	83,72%	89,86%	100%	81,58%	0,934
Bernoulli NB	83,72%	89,86%	100%	81,58%	0,942
Gaussian NB	93,02%	95,89%	100%	92,11%	0,947
SVM	83,72%	89,86%	100%	81,58%	0,884
Logistic Regression	81,40%	88,24%	100%	78,95%	0,900
XGBoost	90,70%	94,44%	100%	89,47%	0,905
Random Forest	90,70%	94,59%	97,22%	92,11%	0,958

Tabela IV: Métricas de desempenho dos modelos - Dados Feminino

Modelo	**Acurácia**	**F1-Score**	**Precisão**	**Recall**	**AUC**
KNN	92,68%	94,92%	96,55%	93,33%	0,921
Bernoulli NB	90,24%	93,33%	93,33%	93,33%	0,952
Gaussian NB	80,49%	86,21%	89,29%	83,33%	0,826
SVM	90,24%	93,33%	93,33%	93,33%	0,924
Logistic Regression	92,68%	94,92%	96,55%	93,33%	0,945
XGBoost	85,37%	90,32%	87,50%	93,33%	0,930
Random Forest	82,93%	88,89%	84,85%	93,33%	0,914

V. Conclusão

Este trabalho teve como objetivo aplicar algoritmos de aprendizado de máquina para auxiliar no diagnóstico de câncer de pulmão, utilizando uma base de dados pública com informações clínicas e laboratoriais, visando combater os problemas de diagnósticos tardios e, consequentemente, melhorar a detecção precoce da doença.

Ao longo do desenvolvimento, foram realizadas etapas fundamentais, como o pré-processamento e o balanceamento dos dados com a técnica SMOTE, a aplicação de sete modelos de classificação e uma análise detalhada do desempenho de cada um por meio de métricas como acurácia, precisão, *recall*, *f1-score* e *AUC*.

Os resultados demonstraram que o modelo *Random Forest* foi o mais eficaz para o problema de classificação de câncer de pulmão neste estudo. Ele apresentou o maior valor de *AUC* (0,991), indicando excelente capacidade de discriminação entre as classes em diferentes limiares de decisão, e obteve *recall* de 100 %, ou seja, conseguiu identificar corretamente todos os pacientes com câncer no conjunto de testes. Essa característica é fundamental em contextos clínicos, nos quais a prioridade é minimizar os falsos negativos — erros que podem levar à não detecção da doença e tratamento tardio.

Embora outros modelos também tenham apresentado bons desempenhos, como o *K-Nearest Neighbors* (KNN) e o *XGBoost*, ambos com alta acurácia (95,18 %) e *AUC* de 0,985, eles não alcançaram o mesmo nível de sensibilidade que o *Random Forest*. O KNN, por exemplo, apresentou precisão perfeita (100 %), mas seu *recall* foi de 94 %, o que implica que alguns casos positivos não foram identificados.

Modelos como *SVM* e Regressão Logística obtiveram *AUC*s de 0,967 e 0,966, respectivamente, confirmando boa discriminação entre as classes, embora com menor acurácia e *recall* em comparação aos líderes. O *Bernoulli Naive Bayes* registrou *AUC* de 0,943, enquanto o *Gaussian Naive Bayes*,

com 0,897, apresentou a menor capacidade discriminativa, refletindo maior suscetibilidade a erros em cenários variáveis.

Em aplicações clínicas, o valor de um falso negativo supera largamente o de um falso positivo; por isso, modelos com *recall* elevado ganham destaque mesmo que incorram em mais alarmes falsos. Que serão verificados e testados por outros exames feitos pelo médicos e outros profissionais da saúde. Nesse sentido, *Random Forest* e *XGBoost*, ambos com *recall* de 100 %, mostram-se particularmente valiosos ao reduzir a chance de casos não detectados.

A análise conjunta de acurácia, *recall*, *f1-score* e *AUC* reforça a necessidade de múltiplas métricas para avaliar modelos em saúde, evitando conclusões baseadas apenas em desempenho médio. A métrica *AUC* permite compreender a separação global entre classes, enquanto o *recall* indica a eficácia do modelo em minimizar falsos negativos, fator crítico no diagnóstico oncológico.

Nesse sentido, a *Random Forest* se destacou como o modelo mais adequado para esta aplicação, oferecendo um equilíbrio robusto entre capacidade discriminativa, sensibilidade máxima e desempenho geral, sendo altamente eficaz na detecção de casos positivos de câncer de pulmão. Isso demonstra que, quando bem configurado e alimentado com dados preparados adequadamente, esse algoritmo pode oferecer suporte confiável para sistemas de apoio à decisão médica.

Como limitações do estudo, destacam-se a utilização de uma base de dados reduzida, com poucos registros e sem informações temporais ou geográficas, o que limita a capacidade de generalização dos resultados para outras populações ou contextos epidemiológicos. Além disso, a ausência de variáveis ambientais impede uma análise mais ampla sobre os fatores externos que podem influenciar a incidência do câncer de pulmão.

Adicionalmente, embora tenha sido aplicada a técnica de balanceamento SMOTE, outras estratégias poderiam ser exploradas para lidar com o desbalanceamento de classes, como o uso de penalizações na função de custo, a adoção de métodos de amostragem alternativos ou o ajuste de limiares de decisão dos classificadores, podem alcançar resultados comparáveis sem a necessidade de aumento artificial dos dados. Conforme discutido por Izbicki (2024) [23], a escolha da técnica de balanceamento influencia diretamente o desempenho e a interpretação dos modelos, isso reforça a importância de avaliar criticamente as estratégias de balanceamento utilizadas, especialmente em cenários sensíveis como o diagnóstico médico, onde a introdução de dados sintéticos pode afetar a interpretabilidade ou a robustez do modelo.

Como trabalhos futuros, sugere-se a aplicação dos modelos desenvolvidos em bases de dados maiores, mais heterogêneas e com recorte temporal, a fim de validar a robustez e a generalização dos resultados em diferentes contextos. Também é recomendada a inclusão de variáveis ambientais relevantes, como qualidade do ar, níveis de poluentes atmosféricos e exposição a agentes tóxicos, o que pode enriquecer a análise preditiva e permitir uma abordagem mais abrangente do problema.

Nesse sentido, pretende-se retomar a proposta inicial deste trabalho, integrando dados ambientais e epidemiológicos do estado de Minas Gerais. Para isso, podem ser utilizados dados públicos do Instituto Brasileiro de Geografia e Estatística (IBGE) e do Instituto Nacional de Meteorologia (INMET) — como indicadores de poluição do ar, dados climáticos e demográficos — juntamente com os registros de internações e mortalidade por câncer de pulmão extraídos do DataSUS. Essa abordagem permitirá uma investigação mais aprofundada da relação entre variáveis ambientais e a incidência da doença em diferentes regiões do estado.

Além disso, recomenda-se a adoção de técnicas de explicabilidade de modelos, como *SHapley Additive Explanations* (*SHAP*) ou *Local Interpretable Model-agnostic Explanations* (*LIME*), com o objetivo de fornecer maior transparência às decisões dos modelos e aumentar a confiança de profissionais da saúde na aplicação prática dessas soluções em ambientes clínicos.

Por fim, este trabalho reforça o potencial da inteligência artificial como ferramenta de apoio ao diagnóstico médico, promovendo decisões mais rápidas, precisas e baseadas em dados.

Referências

[1] Agência Brasil, "Câncer de pulmão no brasil deve crescer 65% até 2040, mostra estudo," 2024. [Online]. Available: https://agenciabrasil.ebc.com.br/radioagencia-nacional/saude/audio/2024-05/cancer-de-pulmao-no-brasil-deve-crescer-65-ate-2040-mostra-estudo

[2] Veja. (2023) Câncer de pulmão: 80% das mortes no brasil têm relação com tabagismo. Acesso em: 15 jul. 2025. [Online]. Available: https://veja.abril.com.br/saude/cancer-de-pulmao-80-das-mortes-no-brasil-tem-relacao-com-tabagismo/

[3] L. H. Araujo, C. Baldotto, G. d. Castro Jr, A. Katz, C. G. Ferreira, C. Mathias, E. Mascarenhas, G. d. L. Lopes, H. Carvalho, J. Tabacof *et al.*, "Lung cancer in brazil," *Jornal Brasileiro de Pneumologia*, vol. 44, no. 1, pp. 55–64, Jan 2018. [Online]. Available: https://doi.org/10.1590/S1806-37562017000000135

[4] Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," *Shanghai archives of psychiatry*, vol. 27, no. 2, p. 130, 2015. [Online]. Available: https://doi.org/10.11919/j.issn.1002-0829.215044

[5] B. Alsinglawi, O. Al-shari, M. Alorjani, O. Mubin, F. Alnajjar, M. Novoa, and O. Darwish, "An explainable machine learning framework for lung cancer hospital length of stay prediction," *Scientific Reports*, vol. 12, 01 2022.

[6] L. Torgo, R. Ribeiro, B. Pfahringer, and P. Branco, "Smote for regression," vol. 8154, 09 2013, pp. 378–389.

[7] J. Wu, X. Zan, L. Gao, J. Zhao, J. Fan, H. Shi, Y. Wan, E. Yu, S. Li, and X. Xie, "A machine learning method for identifying lung cancer based on routine blood indices: Qualitative feasibility study," *JMIR Medical Informatics*, vol. 7, no. 3, p. e13476, 2019. [Online]. Available: https://doi.org/10.2196/13476

[8] Vikas and P. Kaur, "Lung cancer detection using chi-square feature selection and support vector machine algorithm," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 10, no. 3, pp. 2050–2060, 2021. [Online]. Available: http://www.warse.org/IJATCSE/static/pdf/file/ijatcse791032021.pdf

[9] R. P.R., R. A. S. Nair, and V. Gangadharan, "A comparative study of lung cancer detection using machine learning algorithms," in *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 2019, pp. 1–4.

[10] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artificial Intelligence in Medicine*, vol. 23, no. 1, pp. 89–109, 2001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S093336570100077X

[11] M. A. Bhat, “Lung cancer,” 2021, kaggle Dataset. [Online]. Available: https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer
[12] Scikit-learn, “Scikit-learn: Machine learning in python.” [Online]. Available: https://scikit-learn.org/stable/
[13] Matplotlib, “Matplotlib: Visualization with python.” [Online]. Available: https://matplotlib.org
[14] Pandas, “Pandas: Python data analysis library.” [Online]. Available: https://pandas.pydata.org
[15] NumPy, “Numpy.” [Online]. Available: https://numpy.org/pt/
[16] imbalanced learn, “Smote bib documentarion.” [Online]. Available: https://imbalanced-learn.org/stable/index.html
[17] XGBoost, “Xgboost documentation.” [Online]. Available: https://xgboost.readthedocs.io/en/stable/
[18] Y. K. Kumar and R. Priyanka, “Lung cancer identification system to improve the accuracy using novel k nearest neighbour in comparison with logistic regression algorithm,” in *2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*, Jan 2023, pp. 1–5.
[19] A. Al Bataineh, “A comparative analysis of nonlinear machine learning algorithms for breast cancer detection,” *International Journal of Machine Learning and Computing*, vol. 9, no. 3, pp. 248–254, 2019.
[20] A. Bharat, N. Pooja, and R. A. Reddy, “Using machine learning algorithms for breast cancer risk prediction and diagnosis,” in *2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C)*, 2018, pp. 1–4.
[21] S. S. Raoof, M. A. Jabbar, and S. A. Fathima, “Lung cancer prediction using machine learning: A comprehensive approach,” in *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, 2020, pp. 108–115.
[22] M. Mamun, A. Farjana, M. Al Mamun, and M. S. Ahammed, “Lung cancer prediction model using ensemble learning techniques and a systematic review analysis,” in *2022 IEEE World AI IoT Congress (AIIoT)*, 2022, pp. 187–193.
[23] G. O. Assunção, R. Izbicki, and M. O. Prates, “Is augmentation effective in improving prediction in imbalanced datasets?” *Journal of Data Science*, pp. 1–16, 2024.

MINISTÉRIO DA EDUCAÇÃO
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
DEPARTAMENTO DE COMPUTAÇÃO - NG

ATA Nº 32 / 2025 - DECOM (11.56.03)

Nº do Protocolo: 23062.036980/2025-16

Belo Horizonte-MG, 14 de julho de 2025.

Às 9:00 horas do dia 14 do mês de julho de 2025 , realizou-se a sessão pública de defesa do Trabalho de Conclusão de Curso do(a) aluno(a) THALES HENRIQUE BASTOS NEVES, intitulado Análise comparativa de modelos de Machine Learning para detecção precoce de câncer de pulmão, por videoconferência, na plataforma rnp.

Abrindo a sessão, o(a) Presidente da Banca Examinadora de Trabalho de Conclusão de Curso, professor(a) Fábio Rocha da Silva , constituída também pelo servidor Luiz Fernando Pinheiro Ramos (CPA/CEFET-MG)(coorientador), e pelos professores Guilherme Lopes de Oliveira e Luciana Maria de Assis Campos do DECOM-NG , informou aos presentes as regras de condução das atividades da defesa do TCC e passou a palavra ao(à) candidato(a) para apresentação do trabalho. Seguiu-se a arguição pelos examinadores, com a respectiva defesa do(a) candidato(a). Logo após, a banca se reuniu, sem a presença do(a) candidato(a) e do público, para julgamento e expedição do resultado da avaliação, que corresponde a parte da Nota Final da disciplina de TCC, transcrita abaixo:

Item de Avaliação	Nota Orientador(a)/ Coorientador(a)	Nota 1 Banca	Nota 2 Banca	Média das Notas
Desenvolvimento do trabalho durante a disciplina: **30 pontos**.	30	-------	-------	-------
Monografia: formatação, redação e conteúdo do trabalho: **40 pontos**.	37	37	37	37
Apresentação oral: domínio do tema, clareza, tempo: **30 pontos.**	30	30	30	30

Observações e orientações finais da banca:

O resultado da avaliação foi comunicado publicamente ao(à) aluno(a) pela banca, que recebeu as orientações finais. Nada mais havendo a tratar, o(a) Presidente encerrou a sessão e lavrou a presente ATA que será assinada pelos membros participantes da Banca Examinadora.

(Assinado digitalmente em 14/07/2025 11:10)
FABIO ROCHA DA SILVA
PROFESSOR ENS BASICO TECN TECNOLOGICO
DECOM (11.56.03)
Matrícula: 1703246

(Assinado digitalmente em 14/07/2025 11:58)
GUILHERME LOPES DE OLIVEIRA
PROFESSOR ENS BASICO TECN TECNOLOGICO
DECOM (11.56.03)
Matrícula: 3057913

(Assinado digitalmente em 14/07/2025 11:26)
LUCIANA MARIA DE ASSIS CAMPOS
PROFESSOR ENS BASICO TECN TECNOLOGICO
DECOM (11.56.03)
Matrícula: 2557171

(Assinado digitalmente em 14/07/2025 12:00)
LUIZ FERNANDO PINHEIRO RAMOS
ESTATISTICO
CGA (11.72.01)
Matrícula: 1865206

(Não Assinado)
NATALIA COSSE BATISTA
FUNÇÃO INDEFINIDA
DECOM (11.56.03)
Matrícula: 1659511

Visualize o documento original em **https://sig.cefetmg.br/public/documentos/index.jsp** informando seu número: **32**, ano: **2025**, tipo: **ATA**, data de emissão: **14/07/2025** e o código de verificação: **8f0a02f8df**