Câncer de Pulmão: Análise com base nas cacracterísticas do indivíduo da base Lung Cancer do Kaggle

João Paulo Markiewicz
Departamento de Ciência da Computação
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Minas Gerais, Brasil
jmarkiewicz@sga.pucminas.br

Lara Souza
Departamento de Ciência da Computação
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Minas Gerais, Brasil
lara.brigida@sga.pucminas.br

Matheus Sorrentino
Departamento de Ciência da Computação
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Minas Gerais, Brasil
matheus.sorrentino@sga.pucminas.br

Pedro Rodrigues
Departamento de Ciência da Computação
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Minas Gerais, Brasil
pedro.rodrigues.1373336@sga.pucminas.br

Victor Oliveira
Departamento de Ciência da Computação
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Minas Gerais, Brasil
victor.cabral@sga.pucminas.br

ABSTRACT

No artigo "Câncer de Pulmão: Análise com base nas características do indivíduo da base Lung Cancer do Kaggle", foi conduzida uma análise comparativa do desempenho de diferentes algoritmos de aprendizado de máquina na predição do risco de câncer de pulmão em indivíduos fumantes. Utilizamos métricas de remoção de ouliers e dados ausentes, codificação de atributos categóricos, balanceamento com base em oversampling e undersampling, separação em conjunto de treinamento e teste e validação cruzada (cross-validation). Os algoritmos examinados incluíram Random Forest, Decision Tree, Backpropagation e Perceptron. Para avaliar a eficácia dos modelos, foi empregada validação cruzada, revelando que essa abordagem melhorou significativamente o desempenho do modelo. O Random Forest obteve resultados consistentes e ligeiramente superiores em relação aos outros modelos. A Decision Tree teve um desempenho razoável, porém, um pouco inferior em precisão e recall em comparação com o Random Forest. O Backpropagation apresentou o pior desempenho em comparação com os demais métodos. O Perceptron apresentou um desempenho superior ao Backpropagation, mas ainda inferior em relação aos outros dois métodos. Em resumo, os resultados indicam que os algoritmos de aprendizado de máquina podem fornecer insights valiosos para auxiliar na tomada de decisões no contexto do diagnóstico de câncer de pulmão relacionado ao tabagismo. O código utilizado neste projeto está disponível para replicação no GitHub.

Lung Cancer'24, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1234567890

ACM Reference Format:
João Paulo Markiewicz, Lara Souza, Matheus Sorrentino, Pedro Rodrigues, and Victor Oliveira. 2024. Câncer de Pulmão: Análise com base nas cacracterísticas do indivíduo da base Lung Cancer do Kaggle. In *Proceedings of (Lung Cancer'24).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1234567890

1 INTRODUÇÃO

Neste trabalho da disciplina de Inteligência Artificial, foi proposto que deveríamos implementar algoritmos de aprendizado de máquina em uma base de dados de nossa escolha. Desta forma, implementamos o 'Decision Tree', o 'Random Forest', o 'Backpropagation' e o 'Perceptron'. Para isso, utilizamos a plataforma Kaggle e escolhemos a base de dados 'Lung Cancer'. Definimos como objetivo a análise dos atributos para a previsão do atributo "`lung_cancer`", ou seja, para saber se o indivíduo possui câncer de pulmão.

No campo da medicina, a análise de dados desempenha um papel crucial na compreensão e previsão de doenças. Neste estudo, exploramos a análise de dados aplicada ao câncer de pulmão, com o objetivo de desenvolver um modelo preditivo capaz de estimar a probabilidade de um indivíduo desenvolver câncer de pulmão com base em suas características.

Partimos da premissa de que a probabilidade de desenvolver câncer de pulmão é uma medida complexa que depende de múltiplos fatores, como idade, histórico de tabagismo, sintomas respiratórios, histórico de doenças crônicas e outros. Esses fatores têm influência direta na probabilidade de um indivíduo desenvolver câncer de pulmão.

O câncer de pulmão é uma das principais causas de morte por câncer em todo o mundo, sendo responsável por cerca de 1,8 milhão de mortes em 2020, conforme dados da Organização Mundial da Saúde (OMS). Entre os fatores de risco mais significativos estão o tabagismo, a exposição a poluentes ambientais e ocupacionais, e a

predisposição genética. Os sintomas mais comuns incluem tosse persistente, dor no peito, perda de peso inexplicada e falta de ar.

As implicações práticas dessa pesquisa são significativas, não apenas para médicos em busca de estratégias de diagnóstico mais precisas, mas também para pacientes e entusiastas da saúde que buscam uma compreensão mais profunda dos fatores que influenciam o desenvolvimento do câncer de pulmão. Além disso, este estudo pode fornecer insights valiosos para a comunidade médica e científica, ajudando a identificar padrões e tendências no desenvolvimento do câncer de pulmão e orientar intervenções preventivas e de tratamento mais eficazes.

2 DESCRIÇÃO DA BASE DE DADOS

A base de dados "Lung Cancer" fornece informações abrangentes sobre pacientes diagnosticados com câncer de pulmão, incluindo uma variedade de atributos que descrevem suas características e históricos médicos. Na Tabela 1 está uma descrição dos atributos presentes nesta base de dados e suas distribuições, contando com um número total de instâncias igua a 309:

Table 1: Atributos da base

Atributo	Descrição
gender	Gênero do paciente (M / F)
age	Idade do paciente (39 a 87)
smoking	Se é fumante (NO = 1 / YES = 2)
yellow_fingers	Presença de dedos amarelados (NO = 1 / YES = 2)
anxiety	Presença de ansiedade (NO = 1 / YES = 2)
peer_pressure	Influência de pressão dos pares (NO = 1 / YES = 2)
chronic disease	Presença de doença crônica (NO = 1 / YES = 2)
fatigue	Presença de fadiga (NO = 1 / YES = 2)
allergy	Presença de alergia (NO = 1 / YES = 2)
wheezing	Chiado no peito (NO = 1 / YES = 2)
alcohol consuming	Consumo de álcool (NO = 1 / YES = 2)
coughing	Tosse (NO = 1 / YES = 2)
shortness of breath	Falta de ar (NO = 1 / YES = 2)
swallowing difficulty	Dificuldade para engolir (NO = 1 / YES = 2)
chest pain	Dor no peito (NO = 1 / YES = 2)
lung_cancer	Presença de câncer de pulmão (YES/NO)

A base de dados contém informações sobre múltiplas instâncias de pacientes, cada uma registrada com os atributos acima mencionados. O objetivo principal é prever a presença de câncer de pulmão com base nos outros atributos fornecidos. A seguir, descreveremos a quantidade de instâncias e a distribuição dos valores de cada atributo.

Esta base de dados é essencial para o desenvolvimento e a validação de modelos de aprendizado de máquina que possam auxiliar na previsão do câncer de pulmão, proporcionando uma ferramenta valiosa para a comunidade médica na identificação precoce da doença.

3 ETAPAS DE PRÉ-PROCESSAMENTO

O pré-processamento dos dados é uma etapa crucial para garantir a qualidade e a integridade dos dados antes da aplicação de algoritmos de aprendizado de máquina. A seguir, detalhamos todas as etapas de pré-processamento realizadas no conjunto de dados de câncer de pulmão, bem como a justificativa para cada escolha.

3.1 Seleção de colunas

Inicialmente, o conjunto de dados original continha várias colunas. Selecionamos as colunas mais relevantes para caracterizar os pacientes com câncer de pulmão. As colunas selecionadas foram: `"gender"`, `"age"`, `"smoking"`, `"yellow_fingers"`, `"anxiety"`, `"peer_pressure"`, `"chronic disease"`, `"wheezing"`, `"alcohol consuming"`, `"coughing"`, `"shortness of breath"`, `"swallowing difficulty"`, `"chest pain"` e `"lung_cancer"`. A seleção dessas colunas foi baseada em sua relevância clínica e potencial impacto na presença de câncer de pulmão.

3.2 Remoção de atributos com valores ausentes

Identificamos e removemos registros que continham valores ausentes ou nulos em qualquer uma das colunas. A presença de dados faltantes pode comprometer a qualidade das análises subsequentes, por isso optamos por remover esses registros para garantir a consistência dos dados restantes.

3.3 Codificação de atributos categóricos

Para tratar atributos categóricos, realizamos a codificação dos valores da coluna `"gender"` utilizando `"LabelEncoder"`. Essa codificação converteu os valores categóricos em valores numéricos, facilitando a aplicação dos algoritmos de aprendizado de máquina.

3.4 Remoção de Outliers

Utilizamos o método do intervalo interquartil (IQR) para identificar e remover outliers nos dados numéricos. Valores fora do intervalo [Q1 - 1.5 * IQR, Q3 + 1.5 * IQR] foram considerados outliers e removidos para evitar distorções nas análises.

3.5 Balanceamento das classes

O conjunto de dados original estava desequilibrado, com uma maioria de instâncias da classe "YES" (câncer de pulmão), com 270 instâncias, e uma minoria da classe "NO" (sem câncer de pulmão), como 39 instâncias. Para balancear as classes, utilizamos técnicas de oversampling e undersampling:

Oversampling: Aumentamos o número de instâncias da classe minoritária ("NO") utilizando RandomOverSampler.

Undersampling: Reduzimos o número de instâncias da classe majoritária ("YES") utilizando RandomUnderSampler.

Essas técnicas visaram minimizar a perda de informações e maximizar a representatividade das amostras, contribuindo para um modelo mais equilibrado e robusto.

3.6 Separação em conjunto de treinamento e teste

Dividimos os dados em conjuntos de treinamento e teste na proporção de 80/20, utilizando a função `"train_test_split"` com um `"random_state"` de 42 para garantir a reprodutibilidade dos resultados.

3.7 Validação Cruzada

A cross-validation é uma técnica utilizada em aprendizado de máquina e estatística para avaliar a performance de um modelo de forma mais robusta. Ela é especialmente útil para evitar problemas de sobreajuste (overfitting) e para obter uma estimativa mais precisa da capacidade de generalização do modelo.

Utilizamos validação cruzada (cross-validation) para avaliar a performance dos modelos. A validação cruzada foi realizada a partir da etapa 4 (após o balanceamento dos dados), utilizando 5 folds (cv=5) para garantir uma avaliação robusta dos modelos.

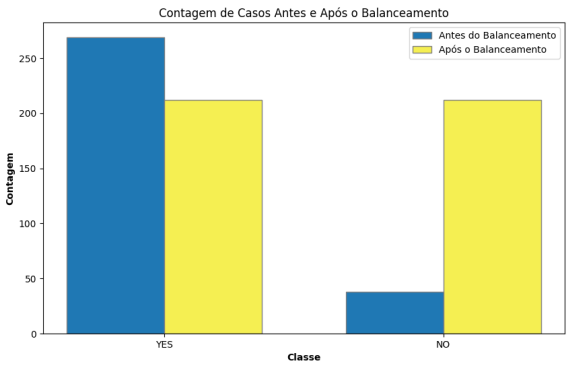

Figure 1: Quantidade de `"lung_cancer"` na base antes e depois do balanceamento

Essas etapas de pré-processamento foram essenciais para garantir a qualidade e confiabilidade das análises realizadas com o conjunto de dados de câncer de pulmão, possibilitando a extração de insights significativos para a detecção e tratamento dessa doença grave.

4 DESCRIÇÃO DOS MÉTODOS UTILIZADOS

Neste projeto, exploramos diversos algoritmos de aprendizado de máquina, incluindo a "Decision Tree" [1], o "Random Forest" [2], "Backpropagation" [3] e "Perceptron" [4] com o propósito de classificar se um indivíduo possui câncer de pulmão com base em suas características.

A "Decision Tree" é uma representação de uma tabela de decisão em forma de árvore, sendo o primeiro algoritmo que utilizamos. Optamos por limitar sua profundidade máxima ('maxdepth=3') para evitar a criação de árvores complexas e pouco generalizáveis.

O "Random Forest" é um algoritmo de aprendizado de máquina que pertence à família de métodos de ensemble, combinando múltiplas árvores de decisão para melhorar a precisão das previsões. Configuramos o modelo com 'maxfeatures' automático para diversificar as árvores e aumentar sua robustez.

O "Backpropagation" é um modelo de aprendizado de máquina inspirado no funcionamento do cérebro humano. Neste projeto, utilizamos uma Rede Neural implementada com o MLPClassifier, com um máximo de 1000 iterações. A rede foi treinada para aprender padrões nos dados e fazer previsões com base neles.

O "Perceptron" é um modelo de aprendizado de máquina inspirado no funcionamento do cérebro humano. Neste projeto, utilizamos uma Rede Neural implementada com o MLPClassifier, com um máximo de 1000 iterações. A rede foi treinada para aprender padrões nos dados e fazer previsões com base neles.

Além disso, foram utilizadas métricas comuns para avaliar a qualidade dos modelos:

Acurácia (Accuracy): A proporção de todas as previsões corretas em relação ao total de previsões feitas.

Precisão (Precision): A proporção de observações positivas previstas corretamente em relação ao total de observações previstas como positivas.

Revocação (Recall ou Sensibilidade): A proporção de observações positivas previstas corretamente em relação ao total de observações positivas reais.

F1-Score: A média harmônica entre precisão e revocação. É útil quando as classes estão desequilibradas.

Essas métricas fornecem diferentes perspectivas sobre o desempenho do modelo e são úteis para avaliar a sua eficácia em diferentes cenários.

Este projeto foi conduzido no ambiente do Visual Studio Code, que oferece uma plataforma integrada para desenvolvimento em Python, simplificando a execução em projetos de aprendizado de máquina.

5 RESULTADOS

Os resultados da análise dos modelos de classificação são apresentados a seguir, com base nas métricas de "precision", "recall" e "F1-Score", bem como na importância dos recursos e na acurácia geral.

Na figura 2 e nas tabelas 2,3,4 e 5 é possível ter uma visão geral sobre o desempenho de cada algoritmo e sobre o que vamos discutir nessa etapa.

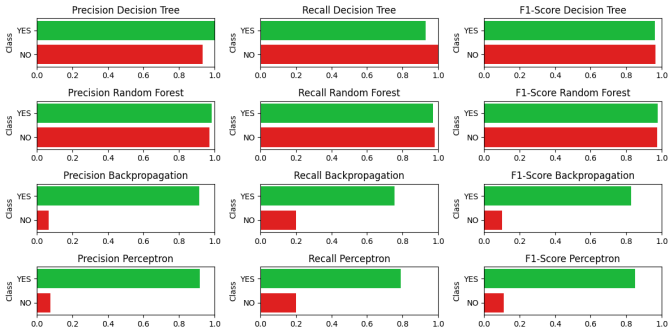

Figure 2: Comparação das métricas para os algoritmos

Table 2: Decision Tree

Metric	YES	NO
Precision	1.000000	0.934463
Recall	0.929236	1.000000
F1-score	0.963068	0.965931
Accuracy	0.964566	

Table 3: Random Forest

Metric	YES	NO
Precision	0.982979	0.972727
Recall	0.971650	0.980952
F1-score	0.976652	0.976100
Accuracy	0.976415	

Table 4: Backpropagation

Metric	YES	NO
Precision	0.914894	0.066667
Recall	0.754386	0.200000
F1-score	0.826923	0.100000
Accuracy	0.709677	

5.1 Análise por classes

Decision Tree: A árvore de decisão apresenta um desempenho razoável, destacando-se na classe "YES", onde a precisão e o recall são relativamente equilibrados. A precisão é perfeita nesta classe, com um valor de 1.000000, enquanto o recall é de 0.929236, resultando em um F1-Score de 0.963068. No entanto, na classe "NO", a precisão é mais baixa (0.934463), embora o recall seja perfeito (1.000000), indicando que o modelo pode ter alguma dificuldade em identificar corretamente instâncias dessa classe. A acurácia geral é de 0.964566.

Random Forest: O Random Forest demonstra um desempenho sólido em ambas as classes, "YES" e "NO", com valores elevados de precisão, recall e F1-Score. Na classe "YES", a precisão é de 0.982979 e o recall de 0.971650, resultando em um F1-Score de 0.976652. Na classe "NO", a precisão é de 0.972727 e o recall de 0.980952, com um F1-Score de 0.976100. A acurácia geral é alta, com um valor de 0.976415, indicando um bom ajuste aos dados.

Table 5: Perceptron

Metric	YES	NO
Precision	0.918367	0.076923
Recall	0.789474	0.200000
F1-score	0.849057	0.111111
Accuracy	0.741935	

Backpropagation: O modelo treinado com backpropagation apresenta um desempenho consideravelmente inferior em comparação com os modelos de árvore de decisão e Random Forest. Na classe "YES", a precisão é de 0.914894 e o recall é de 0.754386, resultando em um F1-Score de 0.826923. No entanto, o desempenho na classe "NO" é significativamente baixo, com uma precisão de apenas 0.066667 e um recall de 0.200000, levando a um F1-Score de 0.100000. A acurácia geral é de 0.709677, indicando que o modelo pode não estar generalizando bem para os dados de teste.

Perceptron: O Perceptron apresenta resultados ligeiramente melhores que o modelo de Backpropagation, mas ainda inferiores aos modelos baseados em árvores. Na classe "YES", a precisão é de 0.918367 e o recall é de 0.789474, resultando em um F1-Score de 0.849057. Na classe "NO", a precisão é de 0.076923 e o recall é de 0.200000, com um F1-Score de 0.111111. A acurácia geral é de 0.741935, o que indica que o Perceptron também pode ter dificuldades em generalizar adequadamente para novos dados.

Em resumo, para o objetivo específico de identificar casos de câncer, o Random Forest seria a melhor escolha. Ele oferece um excelente equilíbrio entre precisão e recall, garantindo que a maioria dos casos de câncer sejam identificados (alta sensibilidade) e que as predições positivas sejam confiáveis (alta precisão). Essa combinação é crucial em aplicações médicas onde tanto a identificação correta quanto a minimização de falsos negativos são vitais.

5.2 Importância dos recursos e regras geradas

A análise da importância dos recursos desempenha um papel crucial na compreensão do processo de classificação em algoritmos de aprendizado de máquina. Para os algoritmos Random Forest, Decision Tree e Rede Neural, os valores de importância dos recursos foram calculados e estão apresentados abaixo:

Decision Tree: [0.043, 0.204, 0.029, 0.225, 0.012, 0.038, 0.132, 0.0, 0.098, 0.00, 0.125, 0.005, 0.089]

Random Forest: [0.042, 0.215, 0.043, 0.085, 0.068, 0.059, 0.074, 0.048, 0.105, 0.066, 0.067, 0.061, 0.067]

Backpropagation: [0.160, 0.096, 0.0070, 0.078, 0.102, 0.156, 0.2224, 0.130, 0.263, 0.208, 0.065, 0.292, 0.138]

Perceptron: [0.094, 0.015, 0.025, 0.292, 0.177, 0.201, 0.307, 0.477, 0.509, 0.315, 0.788, 1.068, 0.377]

Ao observar esses valores, podemos inferir sobre a contribuição de cada atributo no processo de classificação para cada algoritmo. Por exemplo, no Random Forest, atributos como 'AGE' (0.215) e 'ALCOHOL CONSUMING' (0.105) parecem ser os mais importantes.

Além disso, ao analisar as regras geradas pela Decision Tree, podemos entender como as condições específicas dos atributos influenciam na classificação das instâncias:

Regra 1: 'GÊNERO' ≤ 0.5, 'DOENÇA CRÔNICA' ≤ 1.5, 'DEDOS AMARELADOS' ≤ 1.5, 'CONSUMO DE ÁLCOOL' ⇒ Cobertura: 0.015, Classe: NO

Regra 2: 'GÊNERO' > 0.5, 'IDADE' ≤ 65.0, 'FALTA DE AR' ≤ 1.5 ⇒ Cobertura: 0.029, Classe: YES

Regra 3: 'FALTA DE AR' > 1.5 ⇒ Cobertura: 0.75, Classe: NO

Regra 4: 'IDADE' > 65.0 ⇒ Cobertura: 0.333, Classe: YES

Regra 5: 'IDADE' > 67.5 ⇒ Cobertura: 0.333, Classe: NO

Regra 6: 'GÊNERO' ≤ 0.5, 'FALTA DE AR' ≤ 1.5, 'DOENÇA CRÔNICA' > 1.5 ⇒ Cobertura: 0.032, Classe: YES

Regra 7: 'IDADE' > 59.5, 'FALTA DE AR' > 1.5 ⇒ Cobertura: 0.061, Classe: NO

Regra 8: 'IDADE' ≤ 59.5, 'FUMAR' ≤ 1.5, 'TOSSE' ≤ 1.5 ⇒ Cobertura: 0.071, Classe: NO

Regra 9: 'TOSSE' > 1.5 ⇒ Cobertura: 0.5, Classe: YES

Regra 10: 'FUMAR' > 1.5 ⇒ Cobertura: 0.6, Classe: YES

Regra 11: 'IDADE' > 61.5 ⇒ Cobertura: 0.417, Classe: NO

Regra 12: 'GÊNERO' > 0.5 ⇒ Cobertura: 0.176, Classe: YES

Regra 13: 'DEDOS AMARELADOS' > 1.5, 'PRESSÃO DOS COLEGAS' ≤ 1.5, 'DOENÇA CRÔNICA' ≤ 1.5, 'DOR NO PEITO' ≤ 1.5 ⇒ Cobertura: 0.023, Classe: NO

Regra 14: 'IDADE' ≤ 51.0, 'DOR NO PEITO' > 1.5 ⇒ Cobertura: 0.067, Classe: NO

Regra 15: 'IDADE' > 51.0 ⇒ Cobertura: 0.8, Classe: YES

Regra 16: 'DOENÇA CRÔNICA' > 1.5 ⇒ Cobertura: 0.5, Classe: YES

Regra 17: 'PRESSÃO DOS COLEGAS' > 1.5, 'ANSIEDADE' ≤ 1.5, 'FALTA DE AR' ≤ 1.5 ⇒ Cobertura: 0.013, Classe:NO

Regra 18: 'FALTA DE AR' > 1.5 ⇒ Cobertura: 0.8, Classe: YES

Regra 19: 'ANSIEDADE' > 1.5 ⇒ Cobertura: 0.833, Classe: YES

Regra 20: 'GÊNERO' ≤ 0.5, 'FALTA DE AR' ≤ 1.5, 'DIFICULDADE PARA ENGOLIR' ≤ 1.5, 'CONSUMO DE ÁLCOOL' > 1.5 ⇒ Cobertura: 0.019, Classe: YES

Regra 21: 'GÊNERO' > 0.5, 'TOSSE' ≤ 1.5 ⇒ Cobertura: 0.057, Classe: NO

Regra 22: TOSSE' > 1.5, 'DOR NO PEITO' ≤ 1.5 ⇒ Cobertura: 0.25, Classe: YES

Regra 23: 'IDADE' ≤ 62.0, 'DOR NO PEITO' > 1.5 ⇒ Cobertura: 0.063, Classe: YES

Regra 24: 'IDADE' > 62.0, 'CHIADO NO PEITO' ≤ 1.5 ⇒ Cobertura: 0.111, Classe: YES

Regra 25: 'IDADE' ≤ 66.5, 'CHIADO NO PEITO' > 1.5 ⇒ Cobertura: 0.286, Classe: NO

Regra 26: 'IDADE' > 66.5 ⇒ Cobertura: 0.333, Classe: NO

Regra 27: 'IDADE' ≤ 52.0, 'DOR NO PEITO' ≤ 1.5, 'FALTA DE AR' > 1.5 ⇒ Cobertura: 0.007, Classe: YES

Regra 28: 'IDADE' > 52.0 ⇒ Cobertura: 0.5, Classe: NO

Regra 29: 'DOR NO PEITO' > 1.5 ⇒ Cobertura: 0.8, Classe: YES

Regra 30: 'IDADE' > 55.5 ⇒ Cobertura: 0.808, Classe: YES

Regra 31: 'DIFICULDADE PARA ENGOLIR' > 1.5 ⇒ Cobertura: 0.444, Classe: YES

A falta de ar (Regra 3 e Regra 18) e a ansiedade (Regra 19) são fatores predominantes na classificação "YES". Já a idade é um fator significativo, mas pode levar a classificações conflitantes dependendo do limiar específico (Regra 4 vs. Regra 11). No entanto, fumar (Regra 10) e tosse (Regra 9) também são indicadores importantes de classificação "YES". Ademais, regras com coberturas mais baixas fornecem insights adicionais, mas afetam um número menor de casos.

Essas regras destacam diferentes combinações de atributos e seus limiares correspondentes, fornecendo insights sobre como as características individuais dos pacientes influenciam na classificação do câncer de pulmão.

A análise da importância dos recursos e das regras geradas pelos algoritmos oferece uma compreensão mais profunda do processo de classificação, permitindo identificar quais atributos têm maior influência na determinação da classe e como esses atributos interagem

para tomar decisões de classificação. Essas informações são fundamentais para interpretar e melhorar o desempenho dos modelos de aprendizado de máquina na detecção do câncer de pulmão.

5.3 Acurácia Geral

A acurácia geral é uma métrica fundamental para avaliar o desempenho global de um modelo de classificação. Ela representa a proporção de predições corretas em relação ao total de amostras no conjunto de teste. Em outras palavras, a acurácia geral indica a capacidade do modelo de fazer previsões precisas para todas as classes em conjunto.

No contexto do nosso estudo sobre a detecção de câncer de pulmão, a acurácia geral fornece uma medida consolidada da eficácia dos modelos de aprendizado de máquina em identificar corretamente tanto os casos positivos (presença de câncer de pulmão) quanto os casos negativos (ausência de câncer de pulmão). Uma acurácia alta indica que o modelo está fazendo previsões precisas para a maioria das amostras, independentemente da classe.

No entanto, é importante observar que a acurácia geral pode ser enganosa em conjuntos de dados desbalanceados, como é o caso do nosso estudo, onde a classe de câncer de pulmão é minoritária. Nesses casos, a acurácia geral pode ser dominada pela classe majoritária, levando a uma avaliação otimista do desempenho do modelo. Por isso, é crucial complementar a análise da acurácia geral com outras métricas, como precisão, recall e F1-score, especialmente para classes desbalanceadas.

Em resumo, a acurácia geral oferece uma visão geral do desempenho do modelo de classificação, mas deve ser interpretada com cautela, especialmente em conjuntos de dados desbalanceados. Ao combiná-la com outras métricas, podemos obter uma compreensão mais completa do comportamento do modelo em diferentes cenários.

6 CONCLUSÕES

Neste estudo, exploramos a aplicação de diversos algoritmos de aprendizado de máquina na detecção de câncer de pulmão com base em um conjunto de dados desbalanceado. Inicialmente, conduzimos uma análise detalhada da base de dados, aplicando técnicas de pré-processamento para lidar com valores ausentes e outliers, além de balanceamento de classes por meio de oversampling e undersampling.

Durante a etapa de treinamento e avaliação, empregamos quatro modelos de classificação: Decision Tree, Random Forest, Backpropagation e Perceptron. Ao realizar a validação cruzada, notamos melhorias significativas nas métricas de desempenho após o balanceamento das classes, especialmente em termos de precisão e recall para a classe minoritária de câncer de pulmão.

Os resultados finais revelam que o modelo de Random Forest alcançou a maior acurácia geral, seguido de perto pelo modelo Decision Tree. Além disso, a análise feita demonstrou que os modelos foram capazes de identificar corretamente a maioria dos casos de câncer de pulmão, com uma taxa relativamente baixa de falsos positivos e falsos negativos.

No entanto, identificamos oportunidades de melhoria, especialmente na identificação de casos raros e na interpretação dos padrões latentes nos dados. Recomenda-se que estudos futuros explorem técnicas mais avançadas de processamento de dados e modelagem, além de considerar conjuntos de dados mais abrangentes e diversificados.

Em resumo, este trabalho ressalta o potencial dos algoritmos de aprendizado de máquina na detecção precoce de câncer de pulmão e oferece uma base sólida para pesquisas adicionais nesta área crucial da medicina diagnóstica.

7 CÓDIGO DESENVOLVIDO

O código desenvolvido no trabalho pode ser encontrado através do link a seguir: https://github.com/BlackStorm429/IA-Lung_Cancer_-Dataset.

REFERENCES

[1] Datacamp. [n. d.]. *Decision Tree Classification in Python Tutorial.* https://www.datacamp.com/tutorial/decision-tree-classification-python
[2] Datacamp. [n. d.]. *Random Forest Classification with Scikit-Learn.* https://www.datacamp.com/tutorial/random-forests-classifier-python
[3] Machine Learning Mastery. [n. d.]. *How to Code a Neural Network with Backpropagation In Python (from scratch).* https://machinelearningmastery.com/implement-backpropagation-algorithm-scratch-python/
[4] Machine Learning Mastery. [n. d.]. *Perceptron Algorithm for Classification in Python.* https://machinelearningmastery.com/perceptron-algorithm-for-classification-in-python/