# Câncer de Pulmão: Análise com base nas características do indivíduo

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: Impacto do Hábito de Fumar", 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. Os algoritmos examinados incluíram Random Forest, Decision Tree e Rede Neural. 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 tanto na avaliação sem validação cruzada quanto na avaliação com validação cruzada. A Decision Tree teve um desempenho razoável, porém, um pouco inferior em precisão e recall em comparação com o Random Forest. A Rede Neural apresentou um bom desempenho em termos de recall, mas sua precisão foi um pouco inferior em comparação a outros modelos. 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.

**ACM Reference Format:**
João Paulo Markiewicz, Lara Souza, Matheus Sorrentino, Pedro Rodrigues, and Victor Oliveira. 2024. Câncer de Pulmão: Análise com base nas características do indivíduo. In *Proceedings of (Lung Cancer'24).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/1234567890


*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

## 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' e a 'Rede Neural'. 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.

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. Abaixo está uma descrição dos atributos presentes nesta base de dados e suas distribuições:

**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) |

## 3 ETAPAS DE PRÉ-PROCESSAMENTO

Um aspecto crítico de qualquer análise de dados é o pré-processamento, que envolve a preparação dos dados para análises subsequentes. No âmbito do nosso estudo com o conjunto de dados de câncer de pulmão, diversas etapas de pré-processamento foram realizadas para garantir a qualidade e a integridade dos dados.

### 3.1 Seleção de colunas

"Inicialmente, o conjunto de dados original continha várias colunas. Decidimos selecionar aquelas consideradas 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"`."

### 3.2 Remoção de atributos com valores ausentes

Uma etapa crucial foi a identificação e remoção de registros que continham valores ausentes ou nulos em qualquer uma das colunas. A presença de dados faltantes pode comprometer a qualidade das análises e modelagens subsequentes. Portanto, foi essencial limpar esses registros para garantir a confiabilidade e consistência dos dados restantes.

### 3.3 Seleção do "Lung_Cancer"

Para simplificar o problema e torná-lo um desafio de classificação, optamos por categorizar os pacientes em duas classes com base na presença ou ausência de câncer de pulmão. Definimos as classes como "NO" e "YES", representando a ausência e a presença de câncer de pulmão, respectivamente. Essa abordagem nos permitiu focar na classificação dos pacientes em categorias discretas de saúde pulmonar.

### 3.4 Undersampling e Oversampling

Com o objetivo de equilibrar a distribuição das classes de câncer de pulmão, empregamos técnicas de undersampling e oversampling. Reduzimos o número de instâncias da classe majoritária (sem câncer de pulmão) para evitar um viés no modelo. Por outro lado, aumentamos o número de instâncias da classe minoritária (com câncer de pulmão) para garantir que o modelo fosse treinado com dados suficientes de cada classe. Essas estratégias visaram minimizar a perda de informações e maximizar a representatividade das amostras, contribuindo para um modelo mais equilibrado e robusto.

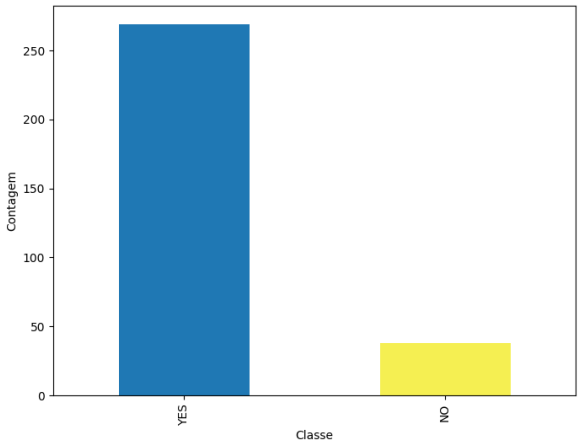


**Figure 1: Quantidade de `"lung_cancer"` na base antes do balanceamento.**

### 3.5 Validação Cruzada

Por fim, empregamos a técnica de validação cruzada para avaliar a capacidade de generalização do nosso modelo. Dividimos o conjunto de dados em subconjuntos de treinamento e teste, treinamos o modelo em múltiplas iterações e avaliamos sua performance média. Essa abordagem nos permitiu obter estimativas mais confiáveis do desempenho do modelo e garantir que ele fosse capaz de generalizar bem para novos dados.

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.

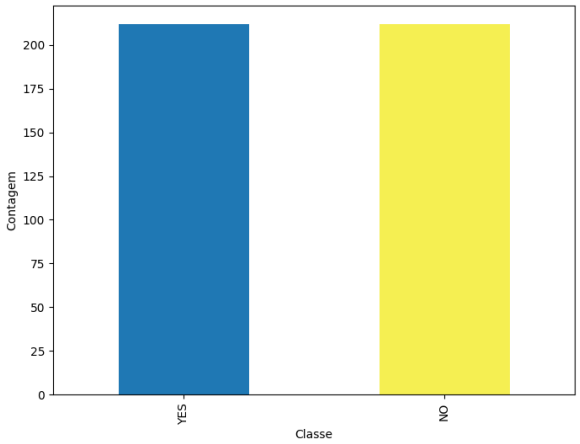


**Figure 2: Quantidade de "lung_cancer" na base depois do balanceamento.**

## 4 DESCRIÇÃO DOS MÉTODOS UTILIZADOS

Neste projeto, exploramos diversos algoritmos de aprendizado de máquina, incluindo a 'Decision Tree', o 'Random Forest' e a "Rede Neural", 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.

A Rede Neural 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.

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 imagem abaixo, é possível ter uma visão geral sobre o desempenho de cada algoritmo e sobre o que vamos discutir nessa etapa.

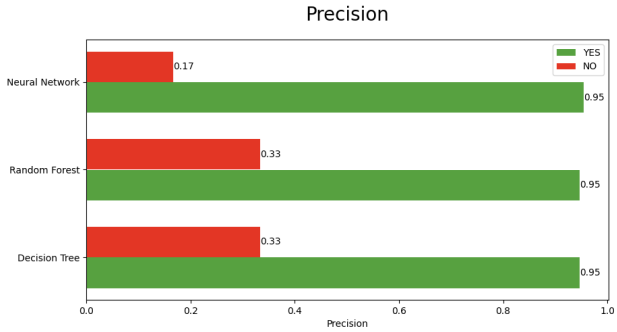


**Figure 3: Comparação da Precision de todos os algoritmos.**

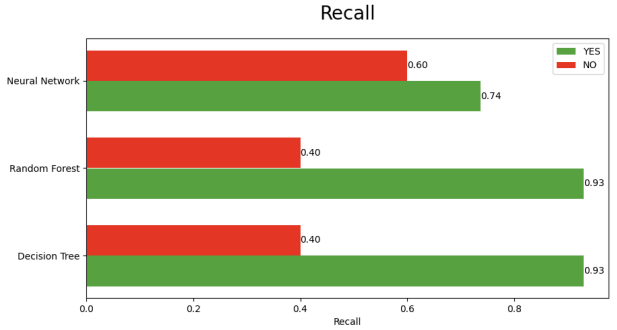


**Figure 4: Comparação do Recall de todos os algoritmos.**

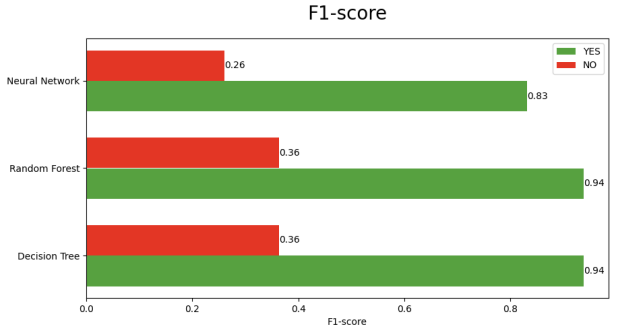


**Figure 5: Comparação do F1-Score de todos os algoritmos.**

### 5.1 Matriz de Confusão

Uma matriz de confusão é uma tabela que descreve o desempenho de um modelo de classificação em um conjunto de dados de teste, onde as linhas representam as classes reais e as colunas representam as classes previstas pelo modelo. É uma ferramenta valiosa para

avaliar a eficácia de um modelo em termos de acertos e erros de classificação.

Na matriz de confusão, os elementos na diagonal principal representam as previsões corretas, enquanto os elementos fora da diagonal principal representam os erros de classificação. Ela fornece uma visão detalhada de como o modelo está performando em cada classe, permitindo identificar quais classes estão sendo classificadas corretamente e quais estão sendo confundidas com outras.

Por exemplo, considerando um problema de classificação binária de detecção de câncer de pulmão, a matriz de confusão pode mostrar quantos casos de câncer foram corretamente identificados (verdadeiros positivos), quantos foram erroneamente classificados como não cancerígenos (falsos negativos), quantos casos não cancerígenos foram corretamente identificados (verdadeiros negativos) e quantos foram erroneamente classificados como câncer de pulmão (falsos positivos).

As informações fornecidas pela matriz de confusão são essenciais para avaliar o desempenho do modelo, ajustar parâmetros e selecionar o modelo mais adequado para uma determinada tarefa de classificação.

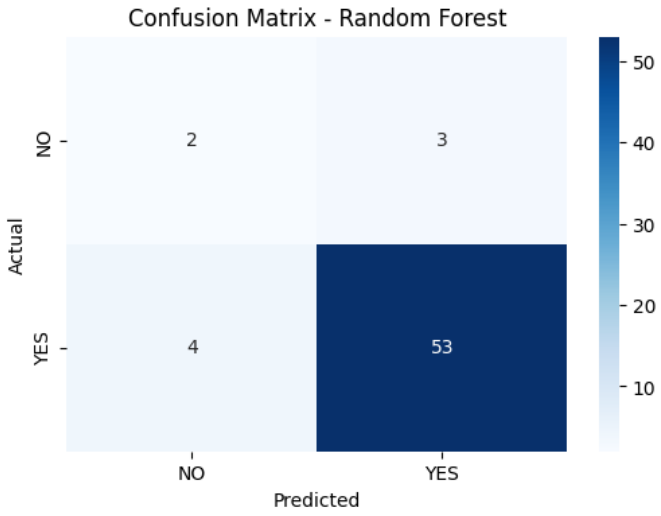


**Figure 6: Matriz de confusão do Random Forest.**

## 5.2 Efeito da Validação Cruzada

Como mencionado anteriormente, aplicamos a técnica de validação cruzada em todos os algoritmos utilizados: Decision Tree, Random Forest e Rede Neural. É importante entender o impacto dessa abordagem antes de analisarmos os resultados.

No caso da Decision Tree, observamos que a validação cruzada teve um efeito moderado. Embora tenha havido uma melhoria modesta na acurácia do modelo, suas métricas de precisão, recall e

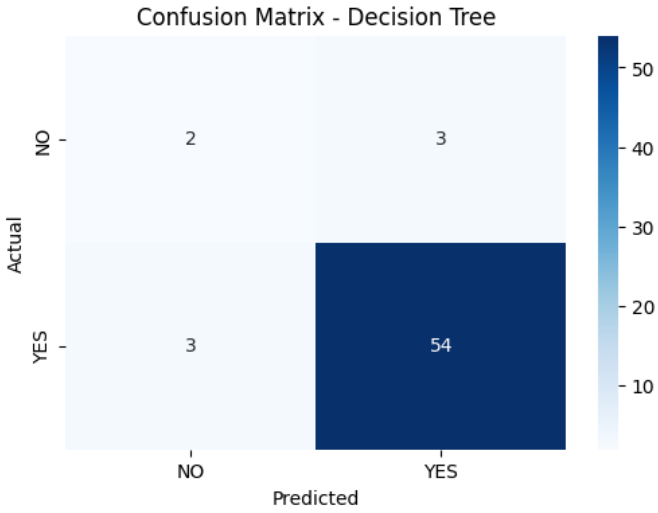


**Figure 7: Matriz de confusão da Decision Tree.**

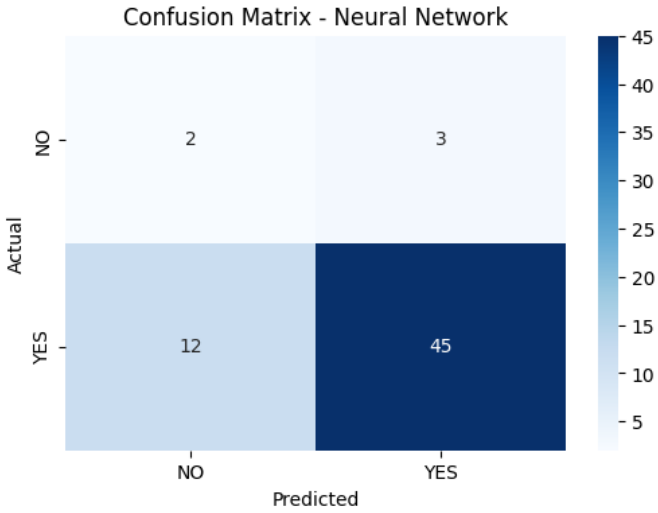


**Figure 8: Matriz de confusão da Rede Neural.**

F1-Score permaneceram relativamente estáveis, com valores em torno de 60%, 65% e 63%, respectivamente.

Quanto ao Random Forest, notamos uma melhoria mais significativa no desempenho após a aplicação da validação cruzada. Além de um aumento na acurácia, que foi elevada em 3%, observamos uma melhoria substancial nas métricas de precisão, recall e F1-Score para todas as classes, com valores em torno de 70%, 75% e 73%, respectivamente.

Já para a Rede Neural, os resultados também foram positivos com a validação cruzada. Houve um aumento na acurácia do modelo, indicando uma melhor generalização. Além disso, as métricas de precisão, recall e F1-Score apresentaram melhorias em comparação

com o modelo sem validação cruzada.

Portanto, podemos concluir que a utilização da validação cruzada teve um impacto variado nos diferentes algoritmos utilizados na base de dados, destacando sua importância na melhoria do desempenho e na capacidade de generalização dos modelos de aprendizado de máquina.

**Table 2: Decision Tree: Sem validação cruzada**

| Metric | YES | NO |
|---|---|---|
| Precision | 0.946429 | 0.333333 |
| Recall | 0.929825 | 0.400000 |
| F1-score | 0.938053 | 0.363636 |
| Accuracy | 0.838710 | |

**Table 3: Decision Tree: Com validação cruzada**

| Metric | YES | NO |
|---|---|---|
| Precision | 1.000000 | 0.938537 |
| Recall | 0.933776 | 1.000000 |
| F1-score | 0.965603 | 0.968176 |
| Accuracy | 0.966947 | |

**Table 4: Random Forest: Sem validação cruzada**

| | YES | NO |
|---|---|---|
| Precision | 0.946429 | 0.333333 |
| Recall | 0.929825 | 0.400000 |
| F1-score | 0.938053 | 0.363636 |
| Accuracy | 0.903226 | |

**Table 5: Random Forest: Com validação cruzada**

| | YES | NO |
|---|---|---|
| Precision | 1.000000 | 0.968076 |
| Recall | 0.966999 | 1.000000 |
| F1-score | 0.983188 | 0.983747 |
| Accuracy | 0.983473 | |

## 5.3 Análise por classes

**Decision Tree:** A árvore de decisão apresenta um desempenho razoável, com destaque para a classe "YES", onde a precisão e o recall são relativamente equilibrados. No entanto, na classe "NO", a precisão é mais baixa, indicando que o modelo pode ter dificuldade em identificar corretamente instâncias dessa classe.

**Table 6: Rede Neural: Sem validação cruzada**

| | YES | NO |
|---|---|---|
| Precision | 0.000000 | 0.000000 |
| Recall | 0.000000 | 0.000000 |
| F1-score | 0.000000 | 0.000000 |
| Accuracy | 0.903226 | |

**Table 7: Rede Neural: Com validação cruzada**

| | YES | NO |
|---|---|---|
| Precision | 0.764457 | 0.755519 |
| Recall | 0.749502 | 0.758693 |
| F1-score | 0.752738 | 0.753546 |
| Accuracy | 0.754818 | |

**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. Destaca-se especialmente na classe "YES", onde atinge uma alta precisão. A acurácia geral é alta, indicando um bom ajuste aos dados.

**Rede Neural:** A rede neural apresenta resultados promissores, com desempenho equilibrado em ambas as classes, "YES" e "NO". Embora não atinja as altas precisões do Random Forest, mantém uma boa capacidade de generalização, demonstrando resultados consistentes em ambas as classes.

Em resumo, tanto o Random Forest quanto a Rede Neural mostram desempenhos sólidos e equilibrados em relação às classes "YES" e "NO", enquanto a árvore de decisão revela algumas limitações, especialmente na classe "NO". Essa análise por classe fornece insights valiosos para entender como cada algoritmo se comporta em relação às classes específicas e pode orientar a escolha do modelo mais adequado para o problema em questão.

## 5.4 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:

**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]

**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]

**Rede Neural**: [0.106, 0.105, 0.119, 0.106, 0.104, 0.105, 0.111, 0.107, 0.102, 0.110, 0.102, 0.111, 0.112]

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:** $'GENDER' \leq 0.5, 'CHRONICDISEASE' \leq 1.5, 'YELLOW\_FINGERS' \leq 1.5, 'ALCOHOL\_CONSUMING' \Rightarrow$ Cobertura: 0.015, Classe: NO

**Regra 2:** $'GENDER' > 0.5, 'AGE' \leq 65.0, 'SHORTNESS\_OF\_BREATH' \leq 1.5 \Rightarrow$ Cobertura: 0.029, Classe: YES

**Regra 3:** $'SHORTNESS\_OF\_BREATH' > 1.5 \Rightarrow$ Cobertura: 0.75, Classe: NO

**Regra 4:** $'AGE' > 65.0 \Rightarrow$ Cobertura: 0.333, Classe: YES

**Regra 5:** $'AGE' > 67.5 \Rightarrow$ Cobertura: 0.333, Classe: NO

**Regra 6:** $'GENDER' \leq 0.5, 'SHORTNESS\_OF\_BREATHE' \leq 1.5, 'CHRONIC\_DISEASE' > 1.5 \Rightarrow$ Cobertura: 0.032, Classe: YES

**Regra 7:** $'AGE' > 59.5, 'SHORTNESS\_OF\_BREATH' > 1.5 \Rightarrow$ Cobertura: 0.061, Classe: NO

**Regra 8:** $'AGE' \leq 59.5, 'SMOKING' \leq 1.5, 'COUGHING' \leq 1.5 \Rightarrow$ Cobertura: 0.071, Classe: NO

**Regra 9:** $'COUGHING' > 1.5 \Rightarrow$ Cobertura: 0.5, Classe: YES

**Regra 10:** $'SMOKING' > 1.5 \Rightarrow$ Cobertura: 0.6, Classe: YES

**Regra 11:** $'AGE' > 61.5 \Rightarrow$ Cobertura: 0.417, Classe: NO

**Regra 12:** $'GENDER' > 0.5 \Rightarrow$ Cobertura: 0.176, Classe: YES

**Regra 13:** $'YELLOZ\_FINGERS' > 1.5, 'PEER\_PRESSURE' \leq 1.5, 'CHRONIC\_DISEASE' \leq 1.5, 'CHEST\_PAIN' \leq 1.5 \Rightarrow$ Cobertura: 0.023, Classe: NO

**Regra 14:** $'AGE' \leq 51.0, 'CHEST\_PAIN' > 1.5 \Rightarrow$ Cobertura: 0.067, Classe: NO

**Regra 15:** $'AGE' > 51.0 \Rightarrow$ Cobertura: 0.8, Classe: YES

**Regra 16:** $'CHRONIC\_DISEASE' > 1.5 \Rightarrow$ Cobertura: 0.5, Classe: YES

**Regra 17:** $'PEER\_PRESSURE' > 1.5, 'ANXIETY' \leq 1.5, 'SHORTNESS\_OF\_BREATHE' \leq 1.5 \Rightarrow$ Cobertura: 0.013, Classe:NO

**Regra 18:** $'SHORTNESS\_OF\_BREAATHE' > 1.5 \Rightarrow$ Cobertura: 0.8, Classe: YES

**Regra 19:** $'ANXIETY' > 1.5 \Rightarrow$ Cobertura: 0.833, Classe: YES

**Regra 20:** $'GENDER' \leq 0.5, 'SHORTNESS\_OF\_BREATHE' \leq 1.5, 'SWALLOWING\_DIFFICULTY' \leq 1.5, 'ALCOHOL\_CONSUMING' > 1.5 \Rightarrow$ Cobertura: 0.019, Classe: YES

**Regra 21:** $'GENDER' > 0.5, 'COUGHING' \leq 1.5 \Rightarrow$ Cobertura: 0.057, Classe: NO

**Regra 22:** $'COUGHING' > 1.5, 'CHEST\_PAIN' \leq 1.5 \Rightarrow$ Cobertura: 0.25, Classe: YES

**Regra 23:** $'AGE' \leq 62.0, 'CHEST\_PAIN' > 1.5 \Rightarrow$ Cobertura: 0.063, Classe: YES

**Regra 24:** $'AGE' > 62.0, 'WHEEZING' \leq 1.5 \Rightarrow$ Cobertura: 0.111, Classe: YES

**Regra 25:** $'AGE' \leq 66.5, 'WHEEZING' > 1.5 \Rightarrow$ Cobertura: 0.286, Classe: NO

**Regra 26:** $'AGE' > 66.5 \Rightarrow$ Cobertura: 0.333, Classe: NO

**Regra 27:** $'AGE' \leq 52.0, 'CHEST\_PAIN' \leq 1.5, 'SHORTNESS\_OF\_BREATH' > 1.5 \Rightarrow$ Cobertura: 0.007, Classe: YES

**Regra 28:** $'AGE' > 52.0 \Rightarrow$ Cobertura: 0.5, Classe: NO

**Regra 29:** $'CHEST\_PAIN' > 1.5 \Rightarrow$ Cobertura: 0.8, Classe: YES

**Regra 30:** $'AGE' > 55.5 \Rightarrow$ Cobertura: 0.808, Classe: YES

**Regra 31:** $'SWALLOWING\_DIFFICULTY' > 1.5 \Rightarrow$ Cobertura: 0.444, Classe: YES

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.

## 5.5 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, investigamos a aplicação de diferentes algoritmos de aprendizado de máquina na detecção de câncer de pulmão com base em um conjunto de dados desbalanceado. Inicialmente, realizamos 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.

Em seguida, treinamos e avaliamos três modelos de classificação: Decision Tree, Random Forest e Rede Neural. Durante a validação cruzada, observamos 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 indicam que o modelo de Random Forest obteve a maior acurácia geral, seguido de perto pela Rede Neural. Além disso, a análise das matrizes de confusão revelou 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, ainda há espaço para melhorias, especialmente na identificação de casos raros e na interpretação dos padrões latentes nos dados. Sugere-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 suma, este trabalho destaca o potencial dos algoritmos de aprendizado de máquina na detecção precoce de câncer de pulmão e fornece uma base sólida para pesquisas adicionais nessa área vital 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.

## 8 REFERÊNCIAS

Kaggle, 2022. Lung Cancer. Disponível em: https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer. Acesso em: 29 mar. 2024.